30/
1.9.1

Psychological Monographs:
General and Applied

Combining the *Applied Psychology Monographs* and the *Archives of Psychology*
with the *Psychological Monographs*

VOL. 74
1960

NORMAN L. MUNN, Editor
Bowdoin College
Brunswick, Maine

Consulting Editors

ANNE ANASTASI
FRANK A. BEACH
ARNOLD M. BINDER
W. J. BROGDEN
ROBERT R. BUSH
JOHN F. DASHIELL
JAMES J. GIBSON
D. O. HEBB
EDNA HEIDBREDER
FRANCIS W. IRWIN
JAMES J. JENKINS
HAROLD E. JONES
DANIEL KATZ
BOYD MCCANDLESS
DONALD W. MACKINNON
QUINN MCNEMAR
LORRIN A. RIGGS
CARL R. ROGERS
RICHARD L. SOLOMON
ROSS STAGNER

ARTHUR C. HOFFMAN, Managing Editor
HELEN ORR, Promotion Manager

Published by
THE AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.
1333 SIXTEENTH STREET N.W., WASHINGTON 6, D.C.

CONTENTS OF VOLUME 74

Whole No.

488 Anoxia, Behavioral Development, and Learning Ability: A Comparative-Experimental Approach. Gilbert W. Meier, Marion E. Bunch, Carson Y. Nolan, and Charles H. Scheidler.

489 Similarities in the Structure of Extreme Social Attitudes. Irving A. Taylor.

490 Perceptual Conditions of Association. Solomon E. Asch, John Ceraso, and Walter Heimer.

491 Meaning and Familiarity in Associative Learning. William Epstein, Irvin Rock, and Carl B. Zuckerman.

492 Relative Effectiveness of Psychotherapeutic Programs: A Multicriteria Comparison of Four Programs for Three Different Patient Groups. G. W. Fairweather, R. Simon, M. E. Gebhard, E. Weingarten, J. L. Holland, R. Sanders, G. B. Stone, and J. E. Reahl.

493 Experimental Studies in the Training of Originality. Irving Maltzman, Seymore Simon, David Raskin, and Leonard Licht.

494 Personality Functions of Symbolic Sexual Arousal to Music. Michael A. Wallach and Carol Greenberg.

495 Opposites Structures, Defenses, and Attitudes. Helen Peak, Barbara Muney, and Margaret Clay.

496 The Effects of Experimentally Controlled Experience upon Recognition Responses. Arnold Binder and Solomon E. Feldman.

497 Re-evaluation of the Meaningfulness of All Possible CVC Trigrams. E. James Archer.

498 The Information Available in Brief Visual Presentations. George Sperling.

499 The Use of Thematic Apperception to Assess Motivation in a Nationwide Interview Study. Joseph Veroff, John W. Atkinson, Sheila C. Feld, and Gerald Gurin.

500 Water Regulation in the Rat: III. The Artificial Control of Thirst with Stomach Loads of Water and Sodium Chloride. Lawrence I. O'Kelly and Robert C. Beck.

501 Evaluating the Educability of the Severely Mentally Retarded Child. Helen Schucman.

502 Sleep Deprivation: Changes in Performance and Physiological Indicants of Activation. Robert B. Malmo and Walter W. Surwillo.

503 Ability Factors and Component Performance Measures as Predictors of Complex Tracking Behavior. James F. Parker, Jr. and Edwin A. Fleishman.

504 Some Variables Affecting Pattern Conception. Eva I. Shipstone.

BUREAU OF EDNL.

Eon..
7. 9. 70
J— 367

Vol. 74, No. 1 Whole No. 488, 1960

Psychological Monographs: General and Applied

ANOXIA, BEHAVIORAL DEVELOPMENT, AND LEARNING ABILITY:

A COMPARATIVE–EXPERIMENTAL APPROACH

GILBERT W. MEIER
Vanderbilt University

MARION E. BUNCH
Washington University, St. Louis

AND

CARSON Y. NOLAN
American Printing House for the Blind

CHARLES H. SCHEIDLER
University of Dayton

Despite the fact that it has been strongly suspected for nearly a century (Little, 1861) that severe anoxia during early development may be responsible for certain neurologic disorders and mental retardation in children who would otherwise be normal, the problem of determining the actual contribution of the anoxia to the abnormality of behavior has remained an exceedingly difficult one. Relatively recent attempts to study the question by controlled animal experimentation in the laboratory have had the advantages of more accurate knowledge and management of the oxygen deficiency to which the animals are exposed, but they have encountered the difficult question of the measurement of psychological processes which may be adversely affected.

The researches discussed in the following monograph have tried to evaluate the effect of oxygen deficiency prior to birth and at birth. They are the culmination of an organized program initiated in the Department of Psychology at Washington University which has been generously supported by grants from the National Institute of Mental Health, United States Public Health Service (M-362 and M-762), and the United Cerebral Palsy, Incorporated (PL-45). Essentially, these studies have shown that a physiological insult, such as anoxia, incurred during early development *can* have an effect on psychological processes, at least those involved in problem solving behavior. The mechanisms by which these changes are effected have not been elucidated. This elucidation is the purpose of further research which, at this time at least, points to a connection with the investigations on the effect of early experience.

The studies reported upon here, along with those which have followed the program initiated at Washington University (Meier, in press; Meier & Menzel, 1955), illustrate a point of view which the authors (among others) feel needs re-emphasis at this time. Within the range of species of animal life adaptable to laboratory conditions, the similiarities between them should prove more important than the differences. Certainly this is true for the understanding of *behavior* rather than behaviors, which are the unique presentations of a particular species.

Clearly, this idea is not new. In those other life sciences in which systems or parts of organisms are of prime interest, e.g., ecology, morphology, and physiology, this truism has been utilized, as well as realized, for some time. For psychology, the usage of a wide range of species offers obstacles of greater proportion than those apparent to the other sciences. Where the time of laboratory maintenance is long, which is the usual case in behavioral research, considerations of size, reproductivity, diet, physiological rhythms, and tractability must be made. In addition, as we have had to do here, the investigator may have to consider the methodology of testing behavior, including the design and complexity of problems available (or which could be feasibly offered), the nature of the growth curve, and the relative developmental stage at birth. All of these appear to be outside the scope of the problem of immediate experimental interest even though the extent to which each or all of them are evaluated will, to some degree, determine the manner of evaluation and extent of generalization of the results and conclusions of the research.

As stated, the impetus for the original researches cited here were the clinical problems centering upon the lack of oxygen during early development. For these investigations, the cat and rat (and, more recently, the chick) were chosen as experimental media. These choices were dictated by the particular phase of the problem studied and the characteristics of the species. (With the recent studies involving the chick, the unusual technical expedience for prenatal deprivation was the major consideration.) We feel that these investigations involving a multispecies approach not only demonstrate the operational practicality of the approach itself, but make the results more tenable, generally, and indicative of behavior changes which are the outcome of a disruption of a sensitive metabolic balance in periods of maximal cellular differentiation and growth.

Some Aspects of the Physiological Mechanisms in Anoxia

A sketch of the principal physiological mechanisms involved in oxygen deprivation should provide some appreciation of the behavior variations discussed in this monograph. The choice of the premature organism delimits the scope of the topic that need concern us here.

The premature organism, in these studies the fetus and the neonate, is characterized by certain metabolic features markedly variant to those found in the mature subject.[1] First, and most obviously, the fetus and the neonate are capable of withstanding degrees of oxygen deficiency far greater than can the more developed animal. For example, rabbit fetuses 29 days after conception survived an atmosphere of pure nitrogen as long as 44 min. Under similar circumstances, the neonate survived 21 min., whereas the 19 day old infant survived only 1½ min. (Glass, Snyder, & Webster, 1944). Depriving chicken embryos and fetuses of oxygen, Meier (unpublished data) noted similar relationships. At the end of 1 day, the median lethal dose (LD_{50}) was 210 min.; at 4 days, 105 min.; at 8 days, 65 min.; at 12 days, 43 min.; at 16 days, 35 min.; and at 19 days, 25 min. In both researches, the ability of the young organism to survive the deficiency was considerably greater than that of the old. The results together suggest a negative exponential function; the ability decreases sharply with maturation.

The explanation of these differences in tolerance to an oxygen lack in the environment lies in the nature of the energy production at the various stages in development. In the early embryo the metabolism is entirely anaerobic; in the adult, the metabolism is largely aerobic. This is a recapitulation of phylogenesis in physiological terms in that with the increasing complexity of structure and function there occurs an increasing efficiency of energy mobilization. With reference to metabolism this has meant a change from anaerobic to aerobic utilization of nutrients in the production of energy; this offers the organism a twelvefold increase in energy output (Himwich, 1951, p. 143).

Second, the limit on organismic survivability to oxygen lack is set, in part, by the peculiarities of the nervous system. The increasingly greater role of aerobic metabolism during development describes this system more aptly than any other in the body. Although at no time in the organism's life is this system completely without anaerobic metabolism, the dominance of aerobic over anaerobic increases rapidly, especially in the period immediately after birth. The arrival at the adult metabolic status depends on the species and its characteristic rate of maturation. In the rat this level is reached at 21 days after birth; in the dog comparable functioning occurs following the lapse of 35 to 40 days (Himwich & Fazekas,

---

[1] A fuller discussion of the relationships between maturation and metabolism will be found in Himwich's *Brain Metabolism and Cerebral Disorders* (1951).

1941). (For this comparison it is assumed that the dog is no more immature at birth than is the rat.)

It follows from the foregoing that primary consideration should be given to the physiological mechanisms of the nervous system. This should not be construed to be a theoretical bias on the part of the authors with regard to the importance of this part of the organism in the totality of its behavior. Insofar as oxygen and the physiology of the body is concerned, the nervous system is the most readily affected, in both the young and the old. When it ceases to function under abnormal conditions of availability of oxygen, eventually, so does the remainder of the organism.

Third, in addition to the uniqueness of the metabolic processes of the nervous system as related to the survivability of the organism, a further fact must be recalled: there are certain delaying actions on the resistance of the adult to physiological insult. This is the fact of the irreversibility of the developed nervous system once the proliferation of the neuronal cells ceases, shortly after birth. Although these cells may continue to grow in size or in the length of their fibrous processes, they do not multiply. Although there occurs a regeneration of parts of the individual cells, regeneration of parts of the system is absent. In terms of the viability and behavior of the organism, this means that in the older individual the debilitation following oxygen deprivation (or some other such insult) will be more extensive and less transitory. Further, artificial aids in the sustenance of life (as some sort of technique for inducing and continuing respiratory activity) will be effective only so long as they are in operation; expectations that the organism will regain the facilities to carry on these vital functions are futile. Lastly, death may follow the deprivation after an appreciable lapse of time because of the inability of the mature system to replace the integral cells which are now fully nonfunctional and degenerated.

Finally, besides the transition in the relative roles of the two forms of metabolism, an equally apparent transition is evidenced by the physiological and behavioral importance given to specific portions of the nervous system at particular phases in development. In terms of the individual's development, the importance of each of the portions follows closely the patterns seen in phyletic growth. As indicated by the rate of differentiation and growth of the constituent cells, the level of oxygen consumption, and the rate of integrated function, the developmental hierarchy proceeds from the lowest centers in the medulla and cord to the highest in the diencephalon and cerebrum. The converse of this is depicted by the adult undergoing oxygen derivation (by drugs, simulated high altitudes, etc.). The cortical functions are the first to cease; the medullary and spinal are the last.

For the birds and mammals the climax in the developmental hierarchy is not reached until after birth. Conversely, in the neonate, centers other than those in the cerebrum are functioning maximally; these are the midbrain, pons, and medulla. As indicated by electroencephalographic activity, a crude measure at best, the first appearance of fully integrated neural activity is in the last third or last quarter of gestation (Jasper, Bridgman, & Carmichael, 1937). At parturition, the activity is still slight, with irregular, slow frequencies and low amplitudes. In most forms the adult pattern is reached considerably later (in the human about 10–12 yrs.) (Charles & Fuller, 1956; F. A. Gibbs & E. L. Gibbs, 1950; Libet, Fazekas, & Himwich, 1941).

So far, the discussion has been upon the descriptive stages in early functional development. Each of these stages encompasses but a brief temporal interval. In such longitudinal researches, as in one reported in this series, these stages, individually, are of unknown significance as determined by the data at hand. That is to say, to date there is insufficient evidence by which one can competently describe the relation between the time of the physiological insult and the particular structures undergoing principal development or the sequelea in the realm of behavior dysfunctions. Possibly, there is a family of curves representing each of the structures, generally alike but each with a different time scale. Each curve rises to a maximum at the period of greatest sensitivity to the insult, then maintains that high point (Levinson, 1952) or, for less complex systems, slowly falls (Ingalls, 1947a, 1947b; Meier, 1953). Physiologically, this is a description of the possible bases of the behavioral phenomena of "critical period" and "psychological irreversibility."

The relation between developmental age (as determined by the level of oxygen consumption) and the severity of effect shows no sign of confounding with specific morphological or functional deviations which could denote impairments in localized structures or systems. This poses an interesting, though still unresolved, problem. Through the stages of fetal and postnatal development, the center making the greatest metabolic demands may not be among those centers most adversely affected. Acute alternations in the young organism are not necessarily followed by appropriate chronic, lasting deviations.

Correlated behavioral–histological studies have born out this observation. Windle, Becker, and Weil (1944) have reported permanent histological sequelae to natal asphyxia as varied, with localized and diffuse degeneration. No relation with particular neural structures was apparent. Severity of deprivation and certain behavioral manifestations likewise appeared to be without concurrence. These animals (guinea pigs) had showed marked sensory–motor dysfunctions (and asphyxia) for the first few days following birth. These included tremors, spasticity, rigidity, somnolence, and paralysis. For the most part, these disappeared within a few days, whereupon the animal assumed the air of normalcy. (The histopathology immediately after deprivation, that is, within one to five days after birth, consisted of, largely, multiple, diffuse capillary hemorrhages.)

Observing only behavior changes, Meier and Menzel (1955) reported that certain changes in behavior did follow deprivation at specific periods in fetal growth. However, rather than use the amount of oxygen metabolized as an index for specified periods of development, they used growth schedules based upon cell size and proliferation. In this experiment, chicken eggs were deprived of oxygen through immersion in water at incubation temperature. Growth schedules were those listed in Hamilton's (1952) *Lillie's Development of the Chick*. On the simple, visually oriented tasks used in the study, that group which had undergone deprivation when the visual–motor system was developing maximally (about the eighth day of incubation) demonstrated the greatest differences in behavior. Like the findings of Windle and Becker (1943), some of the differences disappeared within a week or so after hatching. Unlike their findings, however, differences appeared at this time which were not detected on the first behavioral tests. At this time other groups (those deprived at 12 and 16 days) were different from the controls. As suggested in this report, there may be a differential effect of the early posthatching experience.

Typically, increases in complexity of structure and function and in efficiency of action are related to increases in complexity of behavior need satisfaction and/or to increases in the number of environments compatible to the organism. Growth in any environment means a dependence of the organism upon certain features (moisture, oxygen, pressure, radiation, and temperature) necessary for normal development. Although these of themselves are not sufficient, any limitation in the balance of these features may halt the developmental process and, possibly, life as well. In any environment these features are not constant; to the organism, they vary capriciously. Although these physicochemical processes of development may not be balanced on a pinnacle, certainly they are balanced on a tiny plateau. The confines of this plateau are the limits of adaptability and, ultimately, survivability.

I. NATAL ANOXIA AND LEARNING OF RATS AT MATURITY

Previous studies (Armitage, 1952; Becker & Donnell, 1952; Hurder & Sanders, 1953; Meier & Bunch, 1950; Scheidler, 1953;

Youngdahl, 1948) of the effects of natal anoxia on later learning have produced equivocal results. The present study is one of several directed at a re-examination of this problem. Past studies have generally utilized relatively simple problems to test learning differences among the groups employed. It was believed that (*a*) use of more complex learning tasks, (*b*) measures of retention of such tasks after short time periods, and (*c*) measures of resistance of learned responses to extinction might effectively demonstrate differences unrevealed in the past.

The present study tests the general hypothesis that deprivation of oxygen immediately following birth results in reduced learning ability for rats at maturity. Two levels of deprivation, one-half hour at 2.97% equivalent oxygen (Group I) and one hour at 2.97% equivalent (Group II), were used. The 14 unit Stone type water maze was believed to be a learning task of greater complexity than those previously used. Both groups learned and relearned this problem. Following the learning and relearning of the 14 unit water maze, Group I learned a bar pressing response in the Skinner box which was later extinguished, while Group II learned and reversed a simple position habit.

Subjects

One hundred and fifty-six animals from 18 litters were used in Group I of the study and 59 animals from 9 litters were used in Group II. Of the 156 in Group I, two died in the dessicator under anoxic conditions, and 43 (23 controls, 20 anoxic) died in their home cages before reaching 60 days of age. No significant differences in the mortality rates existed between the groups during this time. Of the 59 in Group II, 10 died in the dessicator under anoxic conditions, and 2 (1 control, 1 anoxic) died later in the home cages. All animals were from the colony of the Department of Psychology, Washington University. They were housed in the racks of metal cages, three to four animals to the cage. The standard colony diet was Purina laboratory chow and water.

Apparatus and Procedure

Deprivation

The apparatus used to produce anoxia in the animals consisted of a vacuum bell jar with a small vacuum pump attached. A manometer measuring from 0 to 90 cm. of mercury was attached to indicate the air pressure within the bell jar. The amount of air flowing through the system was regulated by a stopcock, the adjustment of which allowed airflow through the chamber at all times so as to prevent stagnation of air within the chamber itself.

Through close observation of pregnant females near term, newborn litters were obtained for use within three hrs. of birth. The criterion for rejection of litters was suckling: those litters of animals which showed milk in the stomach were rejected. Acceptable litters were removed from the maternal cage, divided randomly in half, and those animals designated as controls replaced. The experimental animals were marked by excising some of the tissue of the left earbud.

Following marking, they were immediately subjected to an environment of reduced oxygen content. This was accomplished by placing the animals in the bell jar from which the air was gradually withdrawn over a 10 min. period. At the end of this time a difference in pressure of 66 cm. of mercury existed between the inside of the chamber and the outside air. This environment of less than 3% equivalent oxygen was maintained for a period of 30 min. for animals in Group I of the study and 60 min. for animals in Group II. At the end of the appropriate period the pressure was returned to normal; this process also consumed 10 min. Fatalities during deprivation were less than 1% for Group I and 31% for Group II. The animals were returned to their mothers and reared normally. Littermates were later further differentiated by notching the right ear.

Behavior Measures

A straight pathway tank was used to accustom the animals to swimming. This tank was 108 in. long, 15 in. deep, and 8.5 in. wide. The depth of water in the tank was 9 in. At the end of the tank was a wire mesh ladder which led to an enclosed platform.

Animals were given an initial swimming experience between 65 and 72 days of age. This range of age differences was necessary in order to run each litter as a group. The initial swimming experience consisted of five trials per day in the straight runway tank for each of five days.

A 14 unit water T maze of the Stone type was used to measure learning. This apparatus has been previously described by Licklider and Bunch (1946). A wire mesh ladder at the end of the maze led to an enclosed compartment in a revolving horizontal drum where the animals were housed between trials. The temperature of the water in the maze was approximately 70° F. at all times. The water filled the maze to a depth of about 9 in.

Maze training for all groups was started the day following the finish of practice swimming. The animals were run 18 massed trials per day in the multiple T maze until a criterion of one errorless trial was reached. The motivation used was escape from water. Each trial was scored for errors. An error was defined as entrance of the rat's head and shoulders into a blind alley. Time between trials varied depending on the number of animals in the litter and the degree of learning of the animals. Time between trials was never more than 65 min. (in the early stages of learning) or less than 5 min.

Thirty days following the attainment of the criterion for initial learning, the animals were required to relearn the maze. Running conditions were as previously described. The criterion for relearning was again one errorless trial.

Two standard type Skinner boxes were employed with Group I. To each was attached an electric counter which recorded the number of times the bar was pressed. The boxes were adjusted so that pressure on the bar resulted in the sound of the buzzer which was immediately followed by release of a pellet composed of 50% Purina laboratory chow and 50% sugar.

Pretraining for the Skinner box was carried out in a "pseudo" Skinner box. The interior of the "pseudo" box was identical to the Skinner box employed except that it did have a stationary food tray but did not have the bar.

At 150 days of age pretraining for the Skinner box was started for Group I. The animals were placed on a 23 hr. hunger schedule. On each of the 7 consecutive days access to food of the type used for reward in the Skinner box was allowed in the home cage for a one hr. period. Water was available at all times. Three days of training were next given in the "pseudo" Skinner box. During this period the animals were allowed to eat for an hour each day from the food trays of the training box. Their entire daily ration was obtained in this manner.

Following pretraining, the animals were given three days training in the Skinner box. On each of these days the rat was left in the box until 30 rewarded responses had occurred. Extinction of the bar pressing habit was started the day following the last day of the reward training. Extinction of the habit was accomplished by placing the animals in the Skinner box for a 20 min. period each day until a criterion of five or less responses per 20 min. period was reached. During extinction trials no reward was given for bar pressing responses. A record was kept of the responses per period and of the number of periods required for extinction.

A single unit T maze (previously described by Meier & Bunch, 1950) was used to measure habit reversal with Group II of the experiment. This maze was a tank, the stem of which measured 42 in. and the arm 36 in. The maze was 15 in. deep, 8 in. wide in the stem, and 6 in. wide in the arm. The ends of the arms were extended forward 6 in. parallel with the stem, making the ladder by which the animals were permitted to escape invisible from the choice point. Water was maintained at a depth of 9 in. at a temperature of approximately 70° F.

Habit reversal training in the single T maze was started for Group II at 120 days of age. All animals were required through massed practice to learn a simple left turning habit to a criterion of 10 successive errorless trials. Two days following original learning, the animals were required to reverse the left turning habit to a right turning habit. Again learning was by massed practice to a criterion of 10 errorless trials.

Results

Table 1 gives the means and standard deviations for trials and errors for the control and one-half hour anoxic groups for both original learning and relearning in the 14 unit maze. As indicated in the table, none of the mean differences was statistically significant.

TABLE 1

COMPARISON OF NUMBERS OF TRIALS AND ERRORS FOR INITIAL LEARNING AND RELEARNING OF 14 UNIT MAZE BY CONTROL AND ONE-HALF HOUR ANOXIC GROUPS

	Control ($N = 54$)		Anoxic ($N = 57$)			
	Mean	*SD*	Mean	*SD*	*t*	*p*
Trials: Initial Learning	22.89	14.86	22.30	15.24	.20	>.05
Relearning	6.81	4.94	7.42	6.99	.25	>.05
Errors: Initial Learning	173.93	125.36	154.98	104.16	.86	>.05
Relearning	25.69	25.66	25.13	23.62	.12	>.05

TABLE 2

Comparison of Numbers of Responses and Days Necessary for Extinction of a Bar Pressing Habit by Control and One-Half Hour Anoxic Groups

	Control ($N = 15$)		Anoxic ($N = 19$)			
	Mean	*SD*	Mean	*SD*	t	p
Responses	84.66	34.84	97.21	51.88	.79	$>.05$
Days	4.20	1.47	4.47	1.93	.45	$>.05$

Table 2 compares the control and one-half hour anoxic groups for the number of responses and days necessary for extinction of the bar pressing habit in the Skinner box. The lack of significant differences for these variables is demonstrated by the small t ratios given in the table.

Table 3 gives the means and standard deviations for trials and errors for the control and one hour anoxic groups for both original learning and relearning in the 14 unit maze. Examination of the t ratios reported in this table reveals that controls took fewer trials ($p < .01$) than anoxics on both original learning and relearning. Controls also made fewer errors for both sessions ($p < .05$) than did the group with one hour oxygen deprivation.

In Table 4 are found the numbers of trials and errors received by the control and one hour anoxic groups in learning a simple left turning response and then reversing it after a two day period. No significant differences existed between the groups for trials or errors during position learning. Significant differences, favoring the control group, did exist for both trials ($p < .05$) and errors ($p < .01$) when the response was reversed.

A comparison of variances of controls and anoxics of Groups I and II for all variables measured is presented in Table 5. For Group I, significant variance differences occurred only in the case of relearning trials for the Stone maze. All variances differed significantly for Group II except the variances for errors in position learning.

Discussion

The results of this study indicate that deprivation of oxygen for a 30 min. period following birth does not affect later ability of rats to learn and relearn a Stone type maze or to retain a bar pressing habit. Deprivation of oxygen for a 60 min. period

TABLE 3

Comparison of Numbers of Trials and Errors for Initial Learning and Relearning of 14 Unit Maze by Control and One Hour Anoxic Groups

	Control ($N = 23$)		Anoxic ($N = 24$)			
	Mean	*SD*	Mean	*SD*	t	p
Original Learning	20.00	8.96	30.50	13.50	3.09	$<.01$
Trials:						
Relearning	5.56	3.77	10.46	7.64	2.74	$<.01$
Original Learning	149.04	63.48	239.11	160.32	2.50	$<.05$
Errors:						
Relearning	14.83	10.72	40.96	47.17	2.29	$<.05$

TABLE 4

Comparison of Numbers of Trials and Errors for Initial Learning and Reversal of a Position Habit by Control and One Hour Anoxic Groups

	Control ($N = 19$)		Anoxic ($N = 21$)			
	Mean	*SD*	Mean	*SD*	*t*	*p*
Trials:						
Position Learning	12.68	2.18	13.62	5.30	.72	$>.05$
Position Reversal	11.95	1.10	14.52	4.28	2.60	$<.05$
Errors:						
Position Learning	1.47	1.04	1.33	1.25	.38	$>.05$
Position Reversal	3.95	1.82	7.95	3.80	4.21	$<.01$

following birth does affect the later ability of rats to learn and relearn a Stone type maze and to reverse a position habit previously learned. The meaning of the results may best be considered when compared with results of other recent studies.

Windle and Becker (1943) and Becker and Donnell (1952) found differences in the learning abilities of guinea pigs after subjecting halves of litters to severe oxygen deprivation by cutting off intrauterine circulation. Armitage (1952) reported that administration of varying doses of drugs to female rats on the day prior to parturition had a detrimental effect on the later learning ability of the young. Scheidler (1953) found that the offspring of female rats who had been exposed to a simulated altitude of 30,000 ft. for 2 hrs. just prior to birth learned less well than the offspring of normal mothers. Significant differences between anoxics (one-half hour at 3% equivalent oxygen immediately following birth) and controls in ability to reverse previously learned position habits are reported by Meier and Bunch (1950).

In contrast to these positive findings are the negative findings of two studies. Youngdahl (1948) deprived halves of litters of oxygen within 6 hrs. of birth. Deprivation was accomplished by nitrogen dilution of the atmosphere of a chamber to 5% equivalent oxygen for a 25–28 min. period. No differences were found in learning between anoxics and controls at 60 days of age. No differences were found by Hurder and Sanders (1953) when they compared rats which had beeen exposed to a simulated altitude of 34,000 ft. for 34 min. at the age of 1 hr., rats exposed to 30,000 ft. for 3 hrs. at the age of 1 hr., and normal animals.

The authors believe that the deleterious effects of anoxia at birth have been tentatively established. They believe, further, that perhaps two factors are responsible for the varying results of Groups I and II in the present study and the positive and negative results of the studies described above. These factors are (*a*) the maturity of the experimental animals at birth and (*b*) the amount of deprivation received by the experimental groups.

In the studies (Armitage, 1952) in which guinea pigs were used as subjects (*S*s) these factors acting in combination are responsible for the positive results. The guinea pig is very mature at birth with a consequent low resistance to anoxia (Fazekas et al., 1941). Very severe deprivations (almost complete denial of oxygen; in some cases oxygen or gas mixtures were required to revive the experimental animals) were used in these studies. Its severity is attested by the high mortality rate (36%) in the experimental groups. Therefore, low resistance combined with severe deprivation produce the consistent positive findings reported.

It is in the studies of rats that inconsistent findings are reported. These animals being very immature at birth can be expected to have a high degree of resistance to anoxia (Fazekas et al., 1941). Youngdahl (1948), Meier and Bunch (1950), and the experimenters in this study (Group I) used about the same degree of deprivation (approximately 3% equivalent oxygen, or 46,000 ft. for 30 min.). Hurder and Sanders (1952) subjected their animals to 34,000 ft. for 38 min. or to 30,000 ft. for 120 min. Mortalities for the studies reporting them are as follows: Meier and Bunch, overall mortality of the anoxic group exceed the controls by 2.1%; Hurder and Sanders, no losses from anoxic processing, two unspecified animals died; and Group I of the present study, 1% died while being rendered anoxic. Youngdahl does not report these data. With the exception of the limited findings of Meier and Bunch, no differences were found between the experimental and control groups in these studies.

Scheidler (1953) subjected pregnant rats to a simulated altitude of 30,000 ft. for 2 hrs. Armitage (1952) subjected pregnant rats to varying doses of drugs on the day prior to the birth of the young. Mortality rates in these studies were: Scheidler, 39% of the animals died during the deprivation period; and Armitage, 31% overall mortality in the anoxic litters as compared to 16% for the controls. Both of these studies showed positive results.

Consideration of the above findings has led the authors to make the following formulation which is believed to be of heuristic value:

In any large population it is likely that the ability to resist deprivation is normally distributed. Some few individuals will be readily susceptible, some few will be highly resistant, and the majority which lie be-

TABLE 5

Comparison of Variances for Controls and Anoxics of Groups I and II for All Variables

	Control Var.	Anoxic Var.	*F* Ratio	*p*
Group I				
Stone Maze				
Trials:				
Initial Learning	220.82	232.26	1.05	>.05
Relearning	24.40	48.86	2.00	<.01
Errors:				
Initial Learning	15715.13	10849.31	1.45	>.05
Relearning	658.44	557.91	1.18	>.05
df	53	56		
Skinner Box				
Responses to extinction	1213.83	2691.53	2.22	>.05
Days to extinction	2.16	3.72	1.72	>.05
df	14	18		
Group II				
Stone Maze				
Trials:				
Initial Learning	80.28	182.25	2.27	<.05
Relearning	14.21	58.37	4.11	<.01
Errors:				
Initial Learning	4029.71	25702.50	6.38	<.01
Relearning	114.92	2225.01	19.36	<.01
df	22	23		
Single **T** Maze				
Trials:				
Position Learning	4.75	28.09	5.28	<.01
Reversal	1.21	18.32	15.14	<.01
Errors:				
Position Learning	1.08	1.56	1.44	>.05
Reversal	3.31	14.44	4.36	<.01
df	18	20		

tween these extremes will be grouped around some mean resistance value. It is possible to divide the total area of this normal curve into many small areas so that the percentage of cases falling in each area is known. The anoxic tolerances for individuals in each of these small areas can be empirically determined by steadily increasing the amount of deprivation and noting mortalities. It is important to consider what happens to individuals in the small areas adjacent to that area in which fatal deprivation occurred. This is illustrated in Figure 1. Individuals to the right of this small area have all succumbed earlier because their resistance was less. Individuals to the left have been affected to varying extents. A graph can be made giving the proportion of fatalities in the population for each successive increment of deprivation. This assumes a linear relationship between deprivation and proportion of fatalities, although other predictable relationships could be substituted.

It has already been stated (Gomez & Pike, 1909; Grenell, 1946) that a gradient of resistance to anoxia exists; higher mental centers are more susceptible than lower centers. This would imply that a gradient of effect for a given level of deprivation occurs extending in the direction of high resistance. A given deprivation would produce a proportion of fatalities. Individuals falling in that part of the curve adjacent to the area of fatality would be very heavily affected, at the point of death, with such affects diminishing as distance toward high resistance decreases.

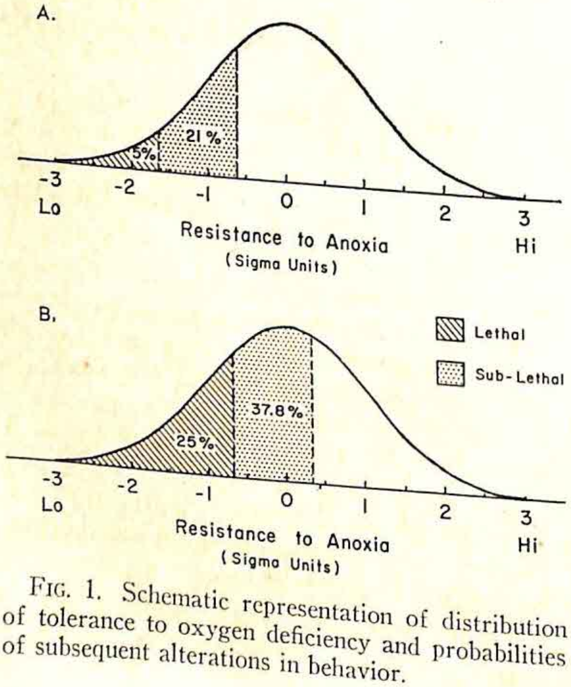


FIG. 1. Schematic representation of distribution of tolerance to oxygen deficiency and probabilities of subsequent alterations in behavior.

The above formulation is important for statistical reasons. Let it be assumed that the normal curve for resistance to anoxia has a mean of zero and a sigma of one. Let it be assumed that the gradient of effect of deprivation extends for one sigma unit to the left of the area of fatality for any given deprivation. If a sample of individuals is submitted to oxygen deprivation to the extent that 5% are fatalities (Figure 1A), what is the effect on the rest of the population? In this case the distance from the edge of the fatal area to the mean of the distribution is 1.64 sigma units. If the gradient of effect extends one sigma unit towards the mean, 21.0% of the total sample has suffered some effect less than death. Let the case where the deprivation is such that fatalities have amounted to 25% of the sample (Figure 1B) be next considered. Here, the distance of the fatal area from the mean is .67 sigma units. If the gradient of effect extends for one sigma unit toward the mean, 37.8% of the sample has been affected to some degree.

In terms of learning scores, this would mean that only part of an experimental group could be expected to have a change in learning ability. The effect of a decrease in learning ability for part of an experimental group would result not only in lowering mean learning scores, but, also, in increasing variability for the experimental group as a whole. This is substantiated by the findings of the present study which are reported in Table 5. In the case of Group II, where the anoxics suffered considerable mortality, the variance for every learning variable save one is significantly greater for the experimentals.

It appears, then, that within certain limits and depending upon the gradient of effect the mortality rate in the experimental sample may be used as an index of the total effect upon that sample. In studies such as those of Youngdahl, Hurder and Sanders, Meier and Bunch, and Group I of the present investigation, where very few or no animals died as a result of deprivation, minimal statistical differences existed between the control and experimental groups because the total effect on the experimental population was small. There probably were, however, individuals in the experimental groups who had been affected. The effect of the scores of these few individuals on the sample statistics was negligible, however.

To summarize the preceding paragraphs: in any species the effects of oxygen deprivation at birth depend upon the maturity of that species at birth and the amount of deprivation. More mature species require less deprivation to produce a given effect. In a sample of any species, the mortality rate of the experimental sample can be taken as an index of the total amount of affect in that sample.

Summary and Conclusions

Since the experimental literature appeared to be equivocal in specifying the role of natal anoxia in the etiology, we conducted this study in which both the severity of oxygen deprivation and type of criterion task were manipulated. One group of rats ($N=57$) was exposed to a simulated altitude of 46,000 ft. (2.91% equivalent oxygen) for 30 min., whereas a second group ($N=24$) was exposed to the same atmospheric conditions for 60 min. Fifty-four littermates served as controls for the first group; 23, for the second.

At the age of 65–72 days, both groups were given pretraining (five trials per day for 5 days) on a straight swimway.

Following the pretraining, the animals learned (to a criterion of one errorless trial) a 14 unit water maze of the Stone pattern. Thirty days after initial mastery, the *S*s relearned the maze.

At the 150th day of age, the animals which had only 30 min. deprivation started learning an instrumental response. This consisted of: (*a*) 7 days adaptation to the food used in a Skinner box; (*b*) 3 days training in a "pseudo" Skinner box; and (*c*) 3 days training in the Skinner box, per se. In the last, the time required for 30 rewarded responses was recorded on each of the days.

Extinction of the bar pressing response in the Skinner box was started after the last of the three days for acquisition. This continued until the animal made five, or less, responses in the daily 20 min. period.

The animals which had 60 min. deprivation learned a position habit in a single unit T maze when 120 days of age. Two days after learning a left turning habit to a criterion of 10 consecutive errorless responses, the *S*s were required to reverse and learn a right turning habit to the same criterion.

From the data gathered in this fashion, we conclude that:

1. Thirty minutes of oxygen deprivation (2.91% equivalent oxygen) had no significant effect on the learning or retention of a multiple unit maze.

2. Further, this degree of oxygen deprivation had no discernible effect on the extinction of a bar pressing habit.

3. Sixty minutes of oxygen deprivation (2.91% equivalent oxygen) produced significant differences in the learning and retention of a multiple unit maze. The deprived animals were significantly inferior in the number of errors made and in the number of trials required to reach criterion on both the learning and relearning of the maze.

4. Although this more stringent degree of oxygen deprivation produced no demonstrable effect on the learning of a simple turning response, it did produce a deleterious effect on the reversal of that habit.

A comparison of these results with the results of similar studies, some of which reported positive findings, suggests that certain variables are of special importance in determining whether positive or negative results will be found. The differences in results among the various studies may be a function of the degree of deprivation used in the study and the physiological maturity of the experimental animals at birth. We have suggested that small amounts of deprivation, while seriously affecting some few individuals in the population, would have little effect upon the statistics derived from a sample of that population. Possibly, the mortality rate in the sample may be validly used as an index of the total number of individuals affected in the sample by a specific amount of deprivation.

II. Delayed Effects of Natal Anoxia on Behavior and Electroencephalographic Activity

Two interests guided the course of the research discussed in this section: (*a*) the aspects of behavior affected by natal oxygen deficiency and (*b*) the relative development of electroencephalographic activity.

The aspects of behavior affected by natal oxygen deficiency. Experimental evidence discussed earlier in this monograph, as well as that published elsewhere (Fazekas et al., 1941) on this problem of natal anoxia, has shown that behavioral dysfunctions may occur, at least on unselected learning tasks. However, a few experimental studies (Becker & Donnell, 1952; Windle & Becker, 1943) and some clinical reports (Despert, 1946; Preston, 1945; Rosenfeld & Bradley, 1948) indicate that selected phases of the total behavior may be more severely impeded than others. Becker and Donnell (1952) report an increased incidence of stereotyped behavior among their experimental *S*s. Despert (1946) has cited instances of extreme levels of anxiety; Preston (1945) and Rosenfeld and Bradley (1948) point out the difficulties of their clinical *S*s in the mastery of tasks which require the use and manipulation of symbols and concepts (e.g., arithmetic).[2]

The *S* used in this research was the cat (a "higher" organism as compared with those used in the other experimentation). We felt that stereotypy and symbol formation could be adequately demonstrated in this *S*. In line with the clinical evidence and that offered by Becker and Donnell, we predicted that those cats which had been deprived of oxygen at birth would evince a lesser variability of behavior and a poorer mastery of symbolic tasks than controls. We predicted poorer learning abilities on those tasks where stereotypy and symbol formation would probably not appear.

Relative development of electroencephalographic activity. Since the pattern of electrocortical activity immediately following anoxemia is well-known (Libet et al., 1941), but little is known regarding delayed or long-term changes, we initiated a program of electroencephalographic recordings on the *S*s used in the behavior analysis. These recordings were scheduled to start at birth and continue at regular intervals for the duration of the experiment.

We predicted a change in the development of the response frequency: in the *S* with natal anoxemia, the development of the frequency of the cortical activity, normally from the slow, irregular waves of the infant to the moderate, regular waves of the adult, would be temporally delayed. At maturity the frequencies for the two groups of *S*s were expected to be the same.

Subjects

In using cats as *S*s, we capitalized upon the existing literature relevant to comparative behaviors on stereotypy, simple discrimination, multiple mazes, and problems evoking symbol formation. Further, by using this *S* we could utilize the considerable amount of material on neuroanatomy, electrocortical activity, and cortical function.

The cats were born of the breeding stock within the colony at Washington University. Although all the young were of known parentage within the colony, the strain was far from homogeneous; it was that which is commonly referred to as "alley cat" variety (a situation which presented some obvious difficulties, especially in the statistical analysis of the results).

Seventy-three litters (272 kittens) were used. These were born into the colony in the interim between March 1, 1950 and April 26, 1952. One hundred and forty-six of the 272 kittens were exposed to a deprivation of oxygen at birth. The remainder, 126 kittens, served as littermate controls.

---

[2] Not all of the clinical studies of this topic have demonstrated positive relationships between neonatal anoxia and intellectual impairment. In a rare predictive study by Apgar, Girdany, McIntosh, and Taylor (1955), no significant correlations were found between blood oxygen content immediately after birth and performance on the Stanford-Binet test administered during early childhood. This failure to find a significant correlation is not easily understood in the light of other findings from the clinic and the laboratory.

Of those exposed to oxygen deprivation, 44 (30.1%) died, during or immediately after exposure. No single factor other than individual differences in tolerance could be denoted in the explanation of these fatalities. A certain percentage, by no means substantial, seemed to have had involvements related to nursing prior to the deprivation.

Behavioral data from the several tasks were collected on 26 of 272 kittens. Of these, 14 were experimental, i.e., they had been deprived of oxygen immediately after birth; the remaining 12 were the littermate controls. Differences in viability between the experimental and control groups did not appear once the deprivation and resuscitation were complete. Some fatalities appeared within the first week after birth (approximately 3% in both groups); these are best described as consequential to maturation level at birth and difficulties incurred during labor and birth processes rather than being causually related to the experimentally induced difficulties. In those litters where early postnatal deaths appeared, both experimental and control groups suffered. The remarkable decrease in the sizes of the groups (about 90%) as seen at birth and at the time of the completion of the experiment is sufficient attestation to the existence of impediments in the rearing of these animals (in large numbers, at least) through the first year of life.

Apparatus and Procedure

Deprivation

The infant cats were exposed to an oxygen deficiency for a single period within the first two hours of postnatal life. From work completed earlier by the authors and from preliminary investigations with infant cats, the severity of the oxygen deprivation was set at 2.91% equivalent oxygen (pressure of 4.16 in. Hg; simulated 46,000 ft.) for 30 min. An additional 15 min. were spent in attaining this condition and another 15 min. returning to normal atmospheric conditions.

The selection of these particular deprivation conditions was originally based on certain limitations in the equipment available when we were working with rats. Contrary to a suspicion engendered early in this study, these conditions did not afford a median lethal dose (LD_{50}) but something less (about LD_{30}), according to the tabulation at the end of the experimentation. (However, continuation with these deprivation conditions permit comparison of data compiled on rats under these conditions.)

In this study, deprivation was achieved through the use of a large, cylindrical decompression chamber (diameter, 54 in.; over-all length, 111 in.; see Figure 2 for details). The decompression was effected by means of a large pump capable of creating a "near vacuum" within a matter of a few minutes. Normally, however, the chamber was evacuated (or filled) at a rate that would be comparable to an ascent (or descent) of 3,000 ft. per min. The conditions within the chamber were indicated on a mercury manometer and a conventional aircraft altimeter. The rate of decompression (and later, return to normal atmospheric pressures) was controlled through a valve sufficient to maintain an equilibrium with the vacuum pump at any level of decompression. This was indicated on an aircraft rate of climb indicator. All instruments were connected directly to the wall of the chamber at a point unaffected by drafts or eddies created by the change of air pressure.

Although the question has been raised repeatedly as to whether the decompression per se has some detrimental effect on the organism, it was felt that the well received work of Paul Bert in 1878 (Van Liere, 1942) furnishes conclusive evidence that it does not. (At least it has no effect within the limits imposed on this experiment.) For this reason the authors are convinced that studies in which decompression is the mode by which deprivation is produced should give results of the same order as in those experiments in which the oxygen is replaced by some inert gas, e.g., nitrogen and CO_2 together. The former method has the advantage of ease and readiness of control; the latter, the advantage of rapidity of effecting the conditions of deprivation.

After this single exposure to this degree of oxygen deprivation, the kittens were returned to their respective mothers and permitted to mature with their littermate controls.

Electroencephalography (EEG)

Shortly before, during, and shortly after the period of deprivation electroencephalographic

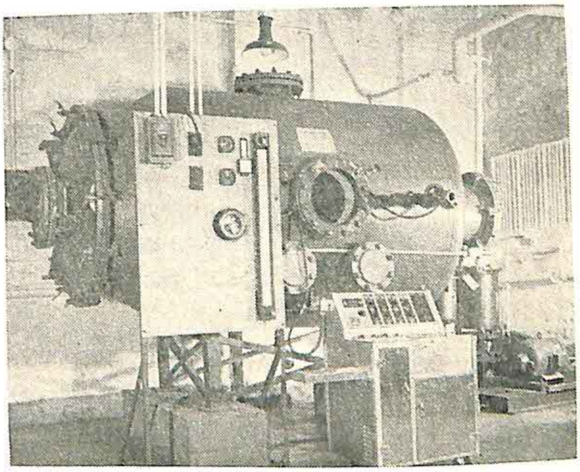


FIG. 2. Photograph of decompression chamber with two-channel electroencephalograph as used for recording during neonatal deprivation.

(EEG) recordings were made on the newly born kittens. As it was impractical to record continuously throughout this period (approximately 1¼ hrs.) samples were made of 2 min. duration for each 10 min. of this entire period.

Follow-up recordings, of about 3 min. duration each, were made the day following the exposure (along with the littermate controls and the other experimental siblings that may not have had recordings made at the time of exposure) and at subsequent monthly intervals through the fourth month. Recordings were again made at the sixth month and at the end of the experiment. At this time they were at least 10 months old.

The monopolar recording technique was used. Leads were placed on the frontal lobes immediately posterior to the orbital ridge and lateral to the median line; on the occipital lobes posterior to an imaginary line extending from the trailing edge of one ear to that of the other, but lateral to the median line; and on both ears (the right ear served as a reference, the left as a ground); six electrodes were used in all (Figure 3). Due to difficulty in the continued placement of the disc electrodes through the movements of the *S*, needle electrodes were employed. (Twenty-seven gauge hypodermic needles were cut ⅛ in. from the base of the stem. The lead wires were soldered to the flat surface of the base remaining.) EEGs were made on a Grass portable two channel recorder (model IIID). For the fourth and sixth month and final EEG recordings, pentobarbital was used to immobilize the animals (dosage: 30 mg./kg. body weight, administered intraperitoneally). Head holders and other restraining devices effectively stabilized the animals for the earlier recordings.[3]

The EEG recordings served as invaluable indices of the vitality of the exposed animals. They correlated well with the overt behavior at this time. When the animals showed a decrease in activity (as the lowest oxygen concentrations were reached), these records showed decrease in frequency and amplitude of the wave forms. Through the 30 min. at this level, both continued to decrease until an extremely low level of electrical activity was attained. In some *S*s, the depression obliterated all recordable electrical activity except at those infrequent times when the infants gasped for air (much less than once each minute).

The overt behavior of the organisms and the concommitant electrical activity evidence the fact that the effective period of deprivation extended beyond the prescribed 30 min. at the low pressure of 66 mm. Hg (2.91% equivalent oxygen). Even and regular breathing was not resumed until some time after the animals were returned to the litter (usually at least an hour).

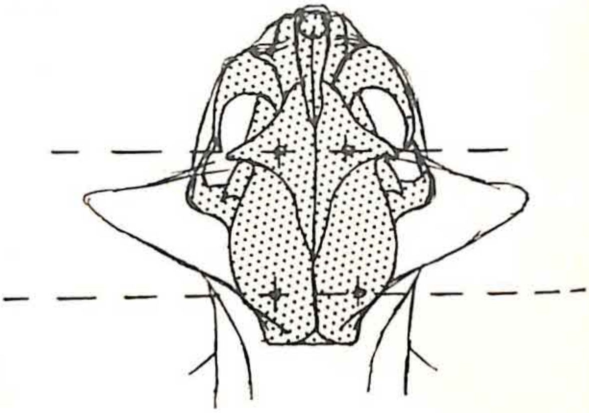


FIG. 3. Representation of head of a kitten to illustrate the placement of electrodes for elecroencephalography with regard to processes on the skull. In addition to the four sites indicated (+), electrodes were also placed on either ear for reference and grounding purposes.

Behavioral Tests

Upon reaching the age of six months, the kittens were started on a battery of behavioral tests (largely learning) which were, in the order of presentation: (*a*) Guthrie puzzle box; (*b*) single unit T maze adapted for black-white discrimination; (*c*) revision of the Hunter Temporal Maze; and (*d*) feline adaptation of the Wisconsin General Test Apparatus. The presentation of all the learning problems had certain aspects in common: The animals were run in blocks of 20 trials each day. At this time the cats were 20–22 hrs. hungry. The reward in all cases consisted of pieces of kidney (about ⅜ in.). Individual animals were trained on the apparatus in small groups in which both the experimental and control *S*s were members.

The various learning tasks will now be described individually:

Guthrie puzzle box. The physical nature of this piece of apparatus was much as it is described in the monograph, *Cats in a puzzle box* by E. R. Guthrie and G. P. Horton (1946). The release mechanism was very similar to that depicted as the "B" variation in this monograph. It consisted of a pole 13 in. in length and ¾ in. in diameter mounted on a hemisphere of wood 4 in. in diameter. Any horizontal movement (at least ¼ in. at the top of the pole) was sufficient to make electrical contact and activate a solenoid to open the door permitting escape and access to the food (see

---

[3] D-Tubocurarine, curare, although superior in theory, had been found to have widely divergent effects on the animals, especially the young kittens. From the self-same dosage insufficient immobilization or fatality resulted. Pentobarbital and other barbiturates were not employed at this time because of a possible synergistic effect to the anoxemia (Himwich, 1951; Snyder, 1949).

Figure 4). The inside dimensions of the retaining box (the main part of the puzzle box) were: 30 in. wide, 24 in. deep (the floor was an extended half circle with a radius of 12 in.), and 34½ in. high. Above the glass face of the box was a blackboard 6½ in. high, situated to indicate the date, the number of the trial, and the number of the *S*.

Within the circuit for the escape mechanism, contacts were made to operate a camera (35 mm.) which "caught" the cat as it manipulated the pole to escape from the apparatus.

The animal was timed from the moment of entry into the retaining box to the time the solenoid was activated and the door opened.

Little preliminary training was required. Five trials were given to familiarize the animal with the goal: a dish of food on the platform in front of the escape box. Ten more trials were given to goal-orient the animal by permitting him to run from the starting compartment, through the retaining box, and out the escape door to the food. (The pole was not in position.)

Five different positions were available for placement of the pole: middle, left front, left rear, right front, and right rear. The first 50 trials were with the pole in the middle position; the next 20 with the pole in the left rear position; the last 10 with the pole in the right front position. Thus, all *S*s had 80 experimental trials.

The reason for the choice of this particular problem stems largely from the article by Windle and Becker (1943) in which the authors alluded to the stereotypy of the anoxic animals in the simple problem used to test their learning ability. Guthrie and Horton claim that by analysis of the photographs taken at the moment of operation of the escape mechanism they were able to infer stereotyped behavior to their *S*s. (By stereotyped behavior Guthrie and Horton meant similarity of bodily position and contact in the act of manipulating the pole.) These two statements were integrated in the construction and usage of the apparatus. Both the experimental and the control *S*s should, according to Guthrie and Horton, show this form of stereotyped behavior. However, if Windle and Becker's statement is valid, at least for cats, the natally asphyxiated group should show this to a greater extent than the control group.

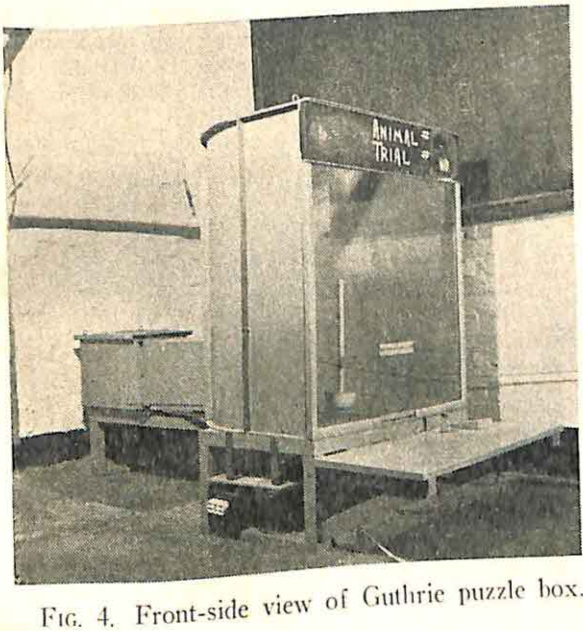


FIG. 4. Front-side view of Guthrie puzzle box.

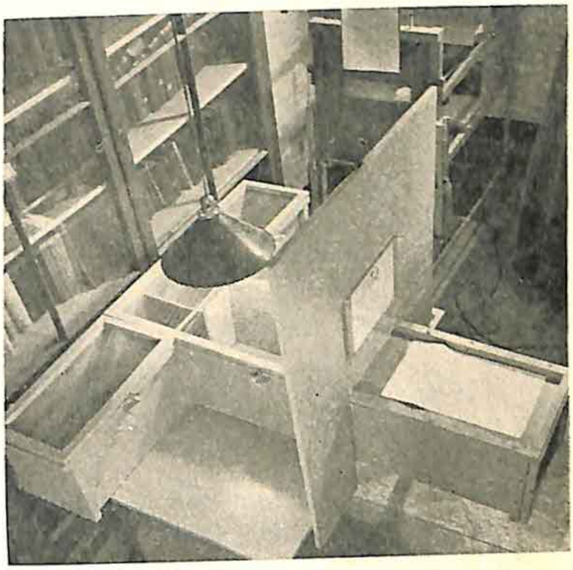


FIG. 5. Top-side view of T maze discrimination apparatus.

As a further check on this issue, the various positions of the pole were utilized as indicated above. That group which possessed the greater degree of stereotypy should operate the escape mechanism in a new, unfamiliar position less readily than formerly.

Discrimination problem: Single unit T *maze.* The apparatus used to test the ability of the cats to discriminate brightness was of conventional design: a simple, enclosed, single unit T maze. The stem of the T was 40 in. long; half of this length comprised the starting box, the other half the straightaway to the choice point. A vertically operated door divided the starting box from the stem. The arms were 24 in. in length. A partition located at the junction of the arms with the stem separated the arms, one from the other. At this point a brass rod was fitted across the stem from which the swinging doors (one-way) were suspended (see Figure 5).

The maze was 10 in. wide in the stem and 8 in. wide in the arms; it was 10 in. deep throughout.

Twenty-five preliminary trials sufficed: 5 to instruct the *S* in running forward from the starting box; 20 to goal-orient him to the left and right goal boxes (swinging doors were not in position at this time).

Subsequent to these 25 trials solid gray doors were placed at the choice point, partly to acquaint the cat with the operation of the doors, partly to teach him left and right position habits (depending on his preference, although instructed that reward was to be found in both goal boxes). After making this choice as to the preferred side (five consecutive trials to one side), the *S* was trained to go to the opposite box in accordance with a noncorrective technique. Here the criterion was 10 consecutive trials to the opposite goal.

Upon mastery of these two position habits the *S* was confronted with the task of learning a brightness discrimination habit. In this instance the gray doors were removed and solid black and white ones substituted. The black door was positive for all *S*s; its position varied randomly from left to right according to a prepared list. The criterion of mastery was 10 consecutive perfect trials. A maximum was arbitrarily set at 300 trials.

Double alternation problem: Hunter temporal maze. The piece of apparatus used in this problem was of an original design suggested by F. H. Palmer (1950). Palmer conjectured that one of the factors that created difficulty in learning Hunter's maze was the number of 90° turns that had to be made which were not relevant to the crucial turns at the choice point. His results on a revision of the maze bore out this contention.

The version of the maze used in this study carried Palmer's plan to the extreme. In Palmer's design the outside corners were rounded leaving still the corners at the intersection *into* the stem leading to the choice point. The current revision consisted of a single T with 10 in. arms and stems. Circular segments (10 in. inside radius) joined the ends of the arms of the T to the stem. Where the two circular segments approached the stem of the T, an arc of the segments was movable so that it could complete the left or the right path (see Figure 6). The passages were 8 in. wide and 10 in. deep throughout.

The starting compartment was below the moveable arc that completed the left and right paths. This was maneuvered into position by depressing a lever, thus raising the arc and the starting compartment (see Figure 7).

The goal box was an extension of the stem of the simple T beyond the choice point.

One-way swinging doors were placed at the ends of the arms of the T and at the start of the stem. These the *S*s operated. A door was also placed at the goal box, concealing it from the animal as he entered the choice point. This the experimenter operated.

The particular double alternation problem was left-left-right-right (LLRR). Reward followed the completion of the entire problem.

At least 90 preliminary trials were required before the *S*s could attempt to learn the pattern. Of these, 20 were spent orienting the animals toward the goal; the animal ran from the starting compartment directly to the goal box with all doors open. The next 10 were the same except that now the swinging door into the stem of the T was closed.

In trials 31–65 the animal learned to make a single choice at the intersection of the T. (The door to the goal box was now closed initially.) The "correctness" of the response, whether left or right,

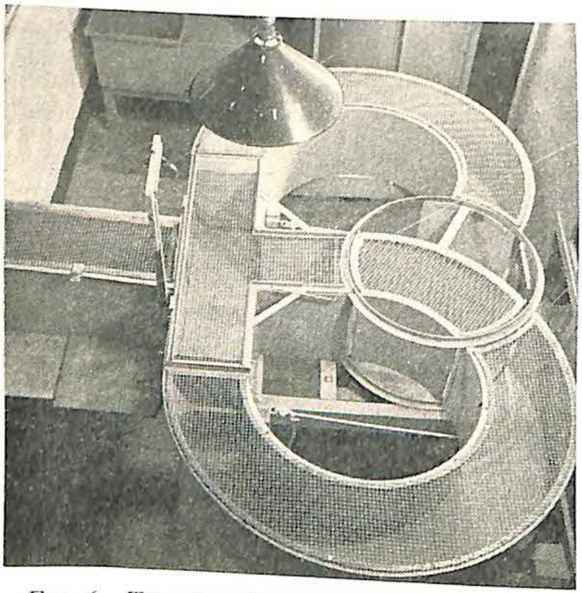


FIG. 6. Top-side view of temporal maze. The starting box is below small segment of arc in the right center of the picture; the goal box is at the left center.

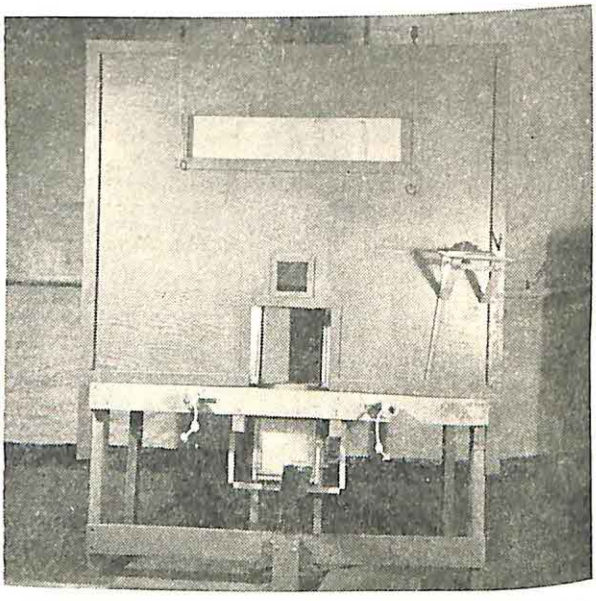


FIG. 7. Front view of temporal maze. The starting box is in the lower center of picture. Depression of treadle raises this box into starting position.

varied randomly according to a prepared list. During the subsequent 20 trials, and in the 5 following them, the *S* learned to put two and three, respectively, of these choices together into a single trial. Again, the correctness of the responses varied in a random fashion. Putting four such choices into a single trial became the actual problem that was to be learned.

The criterion chosen for mastery of the problem was three consecutive errorless trials.[4] Along with the number of trials required to reach this level of performance, the errors made on each of the four legs of the task were recorded.

The choice of the double alternation problem is attributable to the role of symbolic behavior that has been reported for it, even though this reputation has not gone unchallenged (Seward, 1948). It was conceived to serve a position in this battery of tests analogous to those of symbolic functioning used in some of the clinical studies.

Discrimination problem: General test apparatus. The apparatus used here was an adaptation of the general test apparatus designed by H. F. Harlow for his studies of perceptual learning in monkeys (Harlow, 1951). The over-all base dimensions were 20 by 40 in. Of this, the full width (20 in.) and 18½ in. of the length comprised the restraining box from which the cat manipulated the discriminanda. The height of this box was 16 in. A vertically operating door separated the box from the remainder of the apparatus. At this point, inside the vertically operating door, two pieces of plate glass were so arranged as to form a horizontal slit (1¾ in. high) at the level of the tray (3 in.) upon which the discriminanda were placed. This tray extended the full width of the apparatus, being of itself 10 in. wide. Two 2 in. depressions (6 in. between centers) in this tray served as receptacles for the food reward.

A second vertically operating door separated the tray from the experimenter, concealing him from the *S*.

A fluorescent light mounted on the door in front of the *S* illuminated the discriminanda and the tray (see Figure 8).

The cat was given some 20 trials to practice using the paws to secure the food from the tray. In an additional 10 trials he was instructed in the movement of the discriminanda and subsequent exposure of the food.

Thereafter the animal was confronted with the problem of selecting the black object (3½ in. squares were used) over a white object to procure the reward. A corrective type training procedure was used; the *S* persisted until a correct choice was made; this constituted a single trial. As with the previous discrimination task (single unit T maze), 10 consecutive errorless trials constituted the criterion of mastery.

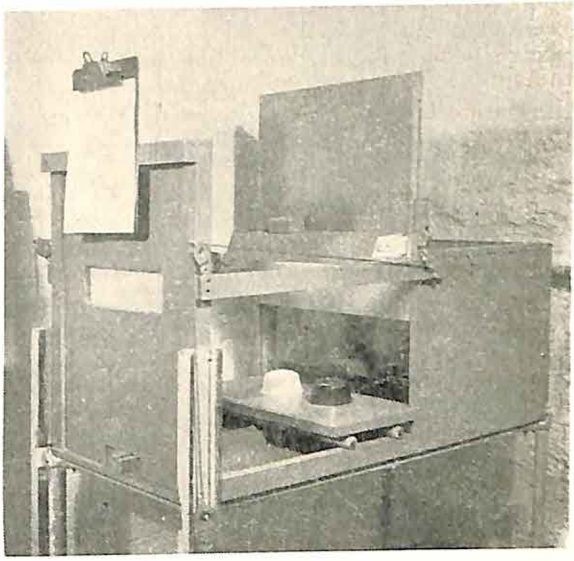


FIG. 8. Front side view of adaptation of Wisconsin General Test Apparatus.

The use of a second discrimination problem was occasioned by several observations of the single unit T maze. Watching the experimental animals in their fixated behavior in this maze, the experimenters felt that part of this behavior could be explained by a differential motivating factor. Possibly the experimental animals were not as highly motivated for food as were their littermate controls. On every trial in the maze the animals were rewarded at least to the extent of being picked up and handled by the experimenter. It was conceivable that this was sufficient to drive the animal forward in the maze, but being the same for both the "right" and "wrong" choices did not reinforce the animal in the learning of the black–white discrimination.

Such an explanation could also be utilized in those instances where the animal did not learn the discrimination but still did not fixate (this included both experimental and control *S*s).

Results

Statistical Analysis

The significance of differences between the two groups of animals in their performances on the behavioral tests was determined by nonparametric procedures. To compare central tendencies, the Mann-Whitney *U* Test (Moses, 1952) was used; for contingencies, Fisher's (1946) exact treatment of 2 by 2 tables was used.

---

[4] The criterion chosen for mastery of the problem was three consecutive errorless trials. If interaction is absent, the probability that this could occur simply by chance is 1:4096.

The preference for the nonparametric techniques over the more traditional parametric forms lies in certain restrictions of the data. Time scores on the Guthrie puzzle box were markedly skewed. Error and trial scores on the single unit T maze and the temporal maze were such that homogeneity of variance was lacking. Admittedly, corrections might have been made for these deficiencies by way of such procedures as suggested by Cochran and Cox (Edwards, 1950; Johnson, 1949) and McNemar (1949). In any event, even with the loss of efficiency inherent in aspects of this analysis of results, the statistics used seemed to us to be the best choice.

As stipulated by the nature of the hypotheses examined here, one-tailed tests of probability prevailed (Jones, 1952).

Subjects

In the final outcome, the number of *S*s (*N*) varied from test to test within the behavioral analysis. Three *S*s died after completing part of the test battery. Some *S*s were withdrawn from individual tests because of a failure to respond. (See the final paragraphs in this section.) Consequently, *N* varied between 21 and 24.

Behavioral Results

Guthrie puzzle box. Two scoring techniques were employed with this apparatus. The first utilized the latencies of response for each trial. Comparisons were made between the two groups as to the mean latency for successive blocks of 10 trials. The trials on three tasks were analyzed: (*a*) the release mechanism in the middle of the restraining cage; (*b*) in the left rear position; and (*c*) in the right front position. In the second technique, the repetitiveness of responses for each animal was determined. The tabulation was made from the photographs taken at the moment of activation of the release mechanism. Only the first 50 trials were considered. (The feeling was maintained that the subsequent 30 trials would show little in the way of stereotypy of response: the rationale for incorporating this instrument in the battery of tests. Reviews of the photographs sustained this belief.)

Significant differences were revealed in the later blocks of trials on the first task and in the first 50 trials, collectively. The performances on the last task (Trials 71–80) also demonstrated statistically significant differences. In every case of significant difference, the latencies for the controls were shorter than for the experimental animals. (Only in one of the nine comparisons made was the median latency for the experimental group shorter than the median for the control group.)

The study of the photographic records proved unrewarding. In this analysis each of the photographs (taken at the time of activation of the release mechanism) was considered and rated on several dimensions of response. To compensate for any dis-

TABLE 6

GUTHRIE PUZZLE BOX

Time Scores

Trials:	1–50		1–10		11–20		21–30		31–40		41–50		51–60		61–70		71–80	
Group:	C*	X**	C*	X**	C*	X**	C*	X**	C*	X**	C*	X**	C*	X**	C*	X**	C*	X**
Median	7.4	10.3	15.7	23.6	8.8	7.9	5.1	9.3	2.9	7.3	2.2	6.2	4.1	7.8	5.4	8.3	6.3	20.5
Mean	9.2	41.6	35.0	116.5	9.6	36.5	6.0	15.6	6.5	19.4	5.0	11.8	7.4	13.8	27.2	15.4	6.3	28.6
SD	3.9	56.8	52.8	141.3	4.3	82.8	3.0	15.5	6.3	26.2	4.7	12.1	5.6	15.8	3.7	15.9	1.2	31.8
p(*U*)	<.05				<.05		<.06		<.05		<.05						<.05	
p(*Run*)																	<.025	

* *N* = 24.
** *N*, the number of *S*s used, varies with each problem. This variance was due to the withdrawals necessitated in each instance.
Note.—Only those probabilities approaching the 1%–5% levels of significance are indicated in this table.

TABLE 7

GUTHRIE PUZZLE BOX

(Percent of responses)

	Response/Total		Part/Total		Side/Total		2 × /Total	
Group:	C	X	C	X	C	X	C	X
Median	.29	.32	.49	.49	.64	.63	.28	.35
Mean	.33	.36	.52	.52	.64	.67	.29	.34
SD	.15	.18	.12	.19	.11	.20	.12	.16

Note.—Probability (*U*)—all greater than .05; *N* = 24.

crepancies in the number of trials available to this analysis, these tabulations were converted to percentages and the following indices were derived: the proportionate frequency of the dominant localized response (the side and part of the body contacting the pole), the frequency of a dominant response not considering the side of the body involved, the frequency of several responses involving one side only, and the frequency of immediate duplications of a localized response. No significant differences were revealed.

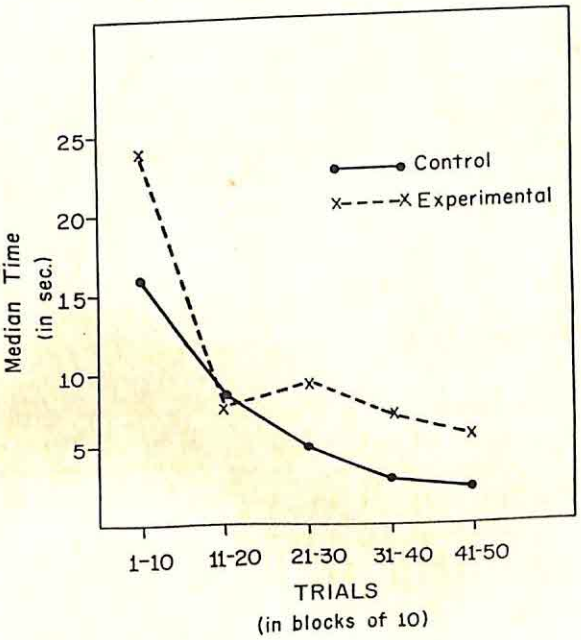


FIG. 9. Comparison of performance, in time scores, of the two groups over the first 50 trials on the Guthrie puzzle box.

Single unit T *maze.* For the three tasks on this apparatus the scoring technique was the same: the number of trials to reach criterion. For the first position habit, the criterion was 5 consecutive responses to one side; for the second position habit and the discrimination task, the criterion was 10 consecutive errorless performances.

Analysis of the performances (in terms of trials to criterion) on the two position tasks by the Mann-Whitney *U* test failed to show significant differences between the two groups.

As there was an upper limit to the number of trials given to each *S*, enforcing an artificial "cut-off" on the distribution of scores, and as there was elicited performance not conducive to learning the discriminatory response, the chi square

TABLE 8

SINGLE UNIT **T** MAZE

(Trials)

	Task		Task	
	1st Position		2nd Position	
Group:	C	X	C	X
Median	20.5	30.0	28.0	28.0
Mean	20.8	40.8	26.2	31.1
SD	6.7	28.3	10.3	11.5
Probability (*U*)	>.05		>.05	

Note.—*N* = 23.

TABLE 9

Frequencies of Successful Performance on Black–White Discrimination Problem

	Learn	Not Learn	Total	*p*
Control	8	2	10	
Experimental	4	8	12	
Total	12	10	22	.034

technique was used with the data on the discrimination problem. Two separate analyses were made: (*a*) the frequencies of those who learned and those who did not learn were compared for the two groups, and the results are given in Table 9; (*b*) the frequencies of those who fixated (as defined earlier) and those who did not are given in Table 10.

TABLE 10

Frequencies of Fixations of Black–White Discrimination Problem

	Fixate	Non-fixate	Total	*p*
Control	0	10	10	
Experimental	5	7	12	
Total	5	17	22	.030

As indicated by the probabilities given to the right of the paradigms, above, both analyses demonstrated significant differences beyond the .05 level. The frequencies of *S*s who did not learn the discrimination and/or fixated were greater for the experimental group than for the control group.

Double alternation problem: Temporal maze. The number of trials necessary to reach the criterion of three consecutive errorless performances and the number of errors made in so doing were the scores used for this problem. The latter, the number of errors made, was recorded in terms of the errors made on each of the four "legs" on a single trial.

When the number of trials were compared, certain significant differences appeared. In terms of the total number of trials to reach the criterion of three consecutive errorless trials, the differences were significant between the .05 and .06 levels, in favor of the control group. The analysis was made with the Mann-Whitney *U* Test. This statistic did not demonstrate a significant difference in the comparison of the numbers of trials required to reach *one* errorless performance. The difference between the groups on the total number of trials seemed to be largely attributable to the differences in the behavior following the completion of the first errorless performance. Statistical significance was indicated

TABLE 11

Temporal Maze

(Trials)

	Score							
	Total Trials		Trials: 0–1		Trials: 1–3		Trials: Perfect	
Group:	C	X	C	X	C	X	C	X
Median	182.0	220.5	97.0	65.5	82.0	159.0	15.0	18.0
Mean	178.0	232.7	86.0	64.7	93.6	167.8	15.7	21.9
SD	47.7	71.2	32.7	46.0	47.9	85.3	6.1	11.2
Probability	<.06		>.05		<.05		>.05	

Note.—$N = 19$.

TABLE 12

TEMPORAL MAZE
(Errors)

	Score									
	Total Errors		Errors 1		Errors 2		Errors 3		Errors 4	
Group:	C	X	C	X	C	X	C	X	C	X
Median	294.0	333.5	69.0	67.5	205.0	227.5	3.0	9.5	4.0	4.5
Mean	274.3	350.8	66.3	64.0	204.7	251.8	6.0	9.3	5.1	5.5
SD	97.6	88.0	30.0	24.4	68.9	86.0	5.0	3.5	4.6	5.6

Probability—all were greater than .05; $N = 19$.

beyond the .05 level in the analysis of this difference.

No significant differences in performance, as measured in errors made, was found. Comparisons were made upon the total number of errors made in reaching criterion as well as upon the errors made on each of the four legs of the task.

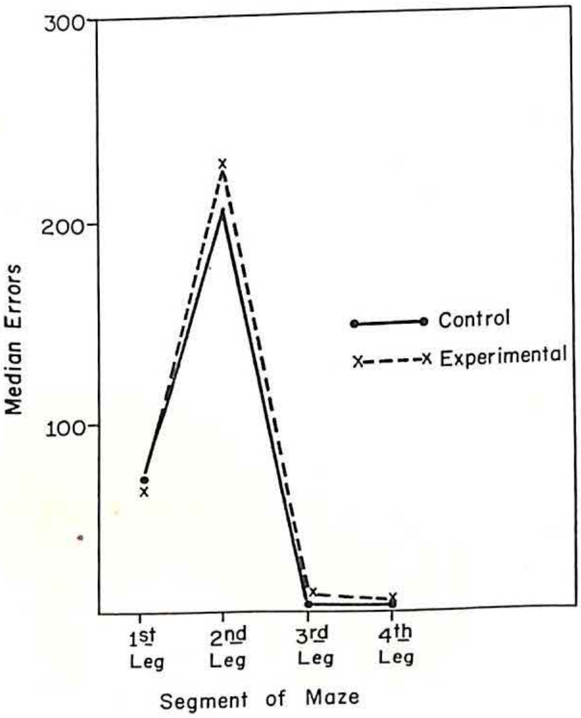


FIG. 10. Comparison of performances, in error scores, of the two groups on each of the four segments ("legs") of the double alternation task.

Discrimination problem: General test apparatus. Although used to corroborate the other test for brightness discrimination (above), the general test apparatus denoted no significant differences between the two groups. There were discrepancies between the medians, in favor of the experimental group, but these lacked statistical confirmation (Mann-Whitney *U* Test).

The results on this apparatus bore low correlation with the similar data from the single unit T maze.[5] All *S*s learned the problem; this included even those who fixated on the earlier task. Some of the most rapid learners were those who failed to learn the correct solution of the discrimination.

In addition to these data compiled from the behavior tests, other impressions of group peculiarities were formed. It was noted, for example, that the experimental animals were more irritable in the learning situations. Should some untoward event occur, which it inevitably did for both groups, it was more likely that these animals would "balk."[6] They would refuse to con-

---

[5] The coefficient of correlation (Kendall's tau) for biserial rank order (1948) with these two sets of data was +.24. On a test of the null hypothesis for the reliability of this correlation, $p = .23$.

[6] These untoward events consisted of unexpected noise within or without the apparatus, unexpected resistance of some movable part, e.g., doors in the temporal maze, or some unprescribed physical restraint or injury.

8-367

TABLE 13

MULTIPLE CHOICE APPARATUS

(Trials)

	Control	Experimental
Median	47.50	44.0
Mean	57.80	45.4
SD	27.9	15.4
Probability	$>.05$	

Note.—$N = 24$.

tinue operation of the problem impelling added encouragement or preliminary type training (even though they were in the midst of the actual learning trials).[7]

Further, the impression prevailed that the experimental animals were less active and responsive in their home cages. The animals tended to remain immobile in a corner of the cage seemingly unmindful of the activity of the other animals (their littermates) or of the proximity of the experimenter and others concerned with the colony.

It was apparent, however, that the single period of oxygen deprivation had little or no effect on normal sexual functioning. Whether it had some effect on the maturation of these functions could not be inferred from the observations. Certainly it did not alter the existence of breeding and maternal behavior.

Electroencephalographic Results

The electroencephalographic data were uniformly negative, i.e., no consistent differences could be found between the two groups *at any stage* of development. Because of the wide variability of the electrical activity between recordings and within a single recording for a given *S*, a statistical analysis of the records on immature *S*s was not feasible. Within the accuracy of the measures, the data on the adults were all but identical, and further statistical analysis was judged unnecessary. In a blind run through a random series of records from both groups, an accurate selection of "controls" and "experimentals" could not be made.[8] (For further data on the electroencephalographic recordings, see Appendix A.)

Discussion

These data indicate that the nature of stereotyped behavior described by Guthrie and Horton (1946) and that described by Windle, Becker, and Donnell (1943, 1952) are not necessarily related, certainly not identical. According to Guthrie and Horton, if one group of *S*s were more stereotyped than the other, then the one would show a greater proportion of related behavioral patterns in the operation of the escape mechanism. It is true that in this study both groups did show the repetitive type of behavior with which Guthrie and Horton concerned themselves; typically, both groups demonstrated this in like amounts.[9]

It follows also, although not necessarily so, that the stereotyped group would show a shorter, or at the very worst, an equal latency of response. The data are quite the contrary. Generally, the experimental group which was hypothetically the more stereotyped was the more sluggish in responding.

The type of stereotypy discussed by Windle and his coworkers was illustrated in the single unit T maze. A given percentage of the *S*s failed to learn the discrimination but selected one arm of the maze and responded persistently to it. Of course, this was an adaptive response, but one demanding less effort and attaining less reward. As only experimental animals demonstrated this type of adaptation, the first hypothesis is given substance.

---

[7] In those instances where this additional acclimation and pretraining were ineffective, the animals were withdrawn from the particular problem.

[8] It is entirely possible that, had there existed a satisfactory means of quantifying data which would depict variations in the voltage input, some conclusion other than "no differences" may have been made of the relative electroencephalographic activity of these two groups.

[9] The series of duplicated responses were not as long, for any *S*, as Guthrie and Horton described with their *S*s.

Two variables, one related to each of these problems (puzzle box and T maze), may bear consideration for the understanding of the differences in these two groups of animals. In the Guthrie puzzle box, the ambiguity of the response–reward relationship is apparent. The control group learned the problem much more rapidly (cf. latencies on individual trials) than did the experimental. It is possible that similarity of successive responses is unimportant to the solution. In the single unit T maze, the delay in reward and the lack of secondary cues are apparent. Again, the control group was superior. Should these two variables be meaningful in the interpretation of the differences between these animals, stereotypy relates to diverse aspects of behavior (or, at least, has two diverse operational definitions).

The performance of the animals on the double alternation problem verifies the notion that the natally anoxic *S*s are inferior in their ability to learn a complicated task involving symbol functioning. Although the groups were not significantly different in the rapidity of solution (i.e., in the number of trials required to make an errorless performance), nor in the pattern of learning the problem (i.e., which leg was learned first, second, etc.), they did differ in the ability to retain that solution and demonstrate that acquisition by consecutive errorless performances. The appearance of the first errorless trial may have been a chance phenomenon ($p = \frac{1}{16} = .0625$). The probability of three consecutive such trials ($p = 1/4096 = .0002$) makes untenable an explanation of these data on the basis of chance alone.

If this double alternation problem is adequate for an operational definition of symbolic behavior, then the data corroborate the observations of Preston (1945) and Rosenfeld and Bradley (1948): Individuals with a history of anoxia during infancy demonstrate a deficit in the ability to acquire and sustain the usage of symbols in their adaptive behavior.

Statistically, the results on the general test apparatus are such that the null hypothesis cannot be rejected. The differences between the two groups of *S*s were not so large (nor in the proper direction as specified by a one-tailed test of significance) as to be significant at the usual levels. Nevertheless, a sizable difference (medians, 8%; means, 21%) did exist suggesting that the experimental animals learned the discrimination *more* readily than did their controls. Whether further experimentation will confirm these differences or the null hypothesis will have little import on the over-all conclusion from these data on the discrimination and stereotypy (as defined in the single unit T maze): The experimental *S*s adapted in a manner not wholly alike that of their controls. The failure to learn a brightness discrimination on the single unit T maze was not effected by an inability to detect differences in brightness discriminanda.

In their entirety the electroencephalographic data were disappointing. This conclusion stems not so much from their failure to confirm or reject one of the hypotheses as from the difficulty of instrumentation and analysis which they engendered. Unusual frequencies appeared in both groups to an equal extent. These were ostensibly consequential to difficulties in the placement of the electrodes (identity of position, in particular), in technical control (electrode resistance, movement of the skin, or presence of underlying muscle groups and sinuses), and in the inability to maintain a quiescent *S* without extensive reference to anesthetics or other drugs. For the younger *S*s, the frequencies were widely variant within a single recording. When repeated recordings were made, this variance was increased. All in all the immature EEGs were remarkably unstable (as judged by the type of analysis used here).

With adult recordings, the frequencies were very much more regular. Frequencies and amplitudes found at one point in the recording were usually found throughout. Repeated recordings on a given *S* were extremely similar. Only at this time did the data appear reliable. At this time no differences between the groups were found—an expected result.

In general, the offhand, unquantified observations on the relative behaviors support a hypothesis which explains the origin of some, maybe all, of the tested behavior dysfunctions in terms of activity levels. This hypothesis states that for the development of normal, adaptive behavior (including functioning in contrivances of the psychologist's sort) a level of responsiveness must be maintained. Especially must this level of responsiveness be maintained during the early stages of development. Physiological insults, as oxygen deprivation, alter this responsiveness, but only temporarily. Cell changes may follow the insult; the cells may degenerate and die but are, shortly, removed and replaced (e.g., cortical neurones may be destroyed by the oxygen deprivation to be replaced, later, by glial cells; the usual failure to find specific, localized, permanent, cerebral histopathology gives empirical support to this hypothesis). Nevertheless, this transient effect of the insult may have far-reaching influences should the period of unresponsiveness be sufficiently long and occur at a critical point in the organism's growth. (Certainly, neither the length of the period nor the point of its occurrence, the critical point, can be specified at this time.)

Surveying the whole of the experimental design and the results gained, we feel that the material, as well as the problems reviewed by Royce (1950), have considerable methodological significance in experimentation of this nature. Using a well-formulated, standardized battery of tests with known intercorrelations denoting conceptualized factors (with operational significance), the experimenter could greatly facilitate the analysis of certain problems of a physiological and/or developmental character (e.g., natal anoxia, prenatal and natal irradiation, hypoglycemia, hypothermia, etc.). Experimental insults, as imposed upon the *S*s in this study, may have a few general or a multitude of particular effects on the organism. To have such a test battery designate these effects and clarify their causal relationships with such variables as age of the *S* and length, frequency, and severity of the insult would alter the isolated existence that such problems as these maintain in the field of physiological psychology, *in toto*. It is plausible that this methodology would extricate from these data such relevancies as would tend to instigate meaningful theorization across the demarcations of specific experimental problems and specific species of animals.

Summary and Conclusions

Behavioral measures were made on 26 cats to evaluate hypotheses concerning the effects of natal anoxemia on learning, stereotypy, symbol formation, and on the development of patterns of electroencephalographic activity. Data of both clinical and experimental origin were presented to indicate the definite possibility of marked effects of anoxemia at birth upon these forms of behavior.

Of these *S*s, 14 were exposed to oxygen deprivation immediately after birth. By means of a decompression chamber the existing atmospheric conditions were reduced to 2.91% equivalent oxygen under which the animals were maintained for 30 min. EEGs were collected at this time and at monthly intervals thereafter. With the exception of the initial recording, concurrent recordings were made on the remaining 12 *S*s, the littermate controls.

At the age of six months the behavior of the animals was observed on a battery of tests: Guthrie puzzle box, single unit T maze, temporal T maze, and a general test apparatus. The cats were to learn a simple brightness discrimination habit on the second and fourth tests and a double alternation habit on the third.

These conclusions were drawn from the observations:

1. The degree of oxygen deprivation incurred in this study altered the *S*s' behavior in such a way as to significantly decrease the rapidity of response in the puzzle box.

2. These animals demonstrated an increased propensity towards poorly adapted responses. They tended to "stereotype" at certain discrimination tasks; rather than learn the proffered discrimination, they would learn a position response.

3. These animals did not manifest differences in ability to learn brightness discriminations in situations in which there was a close spatial and temporal proximity of response to discriminanda and reward.

4. They evinced differences in ability to cope with symbol formation, specifically, in their ability to sustain a symbolic solution after once the solution was gained.

5. They did not show differences in electrical activity (EEG) as seen at adulthood. (Difficulties in instrumentation and analysis prevent conclusion on the development of these patterns.)

III. THE EFFECTS OF PRENATAL ANOXIA ON LEARNING IN RATS

Although the cellular action of an oxygen deprivation prior to birth may be the same as similar deprivation later (Schreiber & Gates, 1938), the differences in the stages of development may be a significant factor circumscribing alternations in the organism's ultimate level of functioning. There is considerable evidence (Breslin, 1956; Ingalls, 1947a, 1947b) that insults at specified periods in the fetal life will effect changes in tissues which are developing maximally at that time. Some evidence (Meier & Menzel, 1955) has appeared indicating that such insults may lead to specific alternations in behavior as well.

In order to appreciate the significance of physiological insults upon the prenatal organism, such as oxygen deprivation, the significance and scope of the developmental activities in this period must be emphasized. This is the time in which occurs most of the growth and development of the organism—more than the sum total of such activities in the remaining vastly more lengthy span of its life. Most of the characteristics for physical and behavioral development are established during this short interval.

The study to be described here was designed to investigate the effects of prenatal oxygen deprivation and, further, to relate the nature of the effect to the age of the prenate at the time of the deprivation. Specifically, the abilities of two groups of animals to learn and retain a simple and a complicated maze pattern were measured. The first group, the experimental, was born to mothers which had been exposed to an oxygen deficiency via a decompression chamber. The second, the controls, was born to mothers of the same stock as the first but which were maintained under normal colony conditions for the entirety of the gestation period.

Part I

Subjects

A population of 80 female albino rats (Wistar strain) from the Buckeye Laboratories in Ohio was the source of the *S*s observed in this study. Randomly, half were designated for the control and half for the experimental groups. All were housed for an appropriate interval with a small group of males of the stock (Wistar strain) in the colony in the Psychology Department at Washington University. The housing of the males with the females was done in such a way that the sires of the experimental offspring were also the sires of the control offspring.

Apparatus and Procedure

Deprivation. Following a positive test for pregnancy (by palpitation), the 40 experimental females were placed in a large decompression chamber and exposed to a simulated altitude of 30,000 ft. (equivalent percentage of oxygen, 6.21; sea level, 20.93) for 120 min. (The pressure in in. of mercury was 8.88 as compared to 29.92 in. at sea level.) The rate of ascent and descent was 3,000 ft. per min. The temperature was maintained at approximately 80° F. at all times during exposure.

Of the 40 animals submitted to the exposure, 12, or 30%, died in the chamber. Based upon observation through the window of chamber, it was thought that all of these animals died within the first 30 min. of exposure. One animal aborted and devoured her young during the first 10 min. of exposure.

Post-mortem examinations revealed that 2 of the 12 animals which died were not pregnant. The post-mortems revealed, further, that the ages of the fetuses at death ranged from an estimated 3 days to an estimated 8 days.

The living animals, following a short recovery period upon removal from the chamber, showed no overt physical disturbances. Moreover, this was true of the 2 deprived animals which died within 5 days and the 3 others which died with 10 days of exposure. Thus a total of 39% of the pregnant females died during deprivation or within 10 days following.

In the control group, 3 of the 40 animals died.

The total number of animals born to anoxic mothers was 180. Of these, 32 died or were destroyed shortly after birth by the mothers. The first 130 animals born were used in the behavior measures and the remaining discarded. To the control mothers, 172 young were born, and of these 35 died or were destroyed by the mothers. The first 130 born were used as the control group in the behavior measures.

The 130 young of the anoxic mothers were born on an average of 7 days following exposure, with the extremes being 2 and 11 days. Since the action of the oxygen deprivation upon the duration of gestation is not clear, no inferences can be drawn as to the exact fetal age at the time of exposure.

Behavior measures. The test situation utilized in measuring the animals' behavior included a straight water tank (a straight "swimway"), a fourteen unit T water maze, and a single unit T water maze.

The straight swimway served a dual purpose: to pretrain the animals for the mazes and to discern swimming speeds of both the anoxic and control groups. The tank used was 108 in. long, 8.5 in. wide, and 15 in. deep. The depth of the water in the tank was 9 in.

The animals were placed in a box suspended over the water at the starting end of the tank and then dropped 3.5 in. to the water through the hinged bottom of the box. The animals then swam to the ladder and goal platform (and warmth) at the opposite end of the tank. Each rat had 10 trials in this task on the 59th day after birth.

On their 60th day of age, the animals of both the control and the experimental groups were weighed. One hundred of each group were started on the multiple T maze.

The maze used was a 14 unit T maze (modified Stone pattern) filled with water to a depth of 9 in. In the scoring procedures, three types of errors were recognized and recorded, along with time and trials to criterion. The three types of errors scored were as follows: retracing in a section of the correct pathway, entering a wrong pathway, and entering either end of a cul-de-sac. Duplicated errors were individually recorded. Five gates were located at points as specified in an earlier report from this laboratory (Licklider & Bunch, 1946).

Each animal was given three trials a day until it reached a criterion of one errorless trial. Any animal spending more than 30 min. in the water in any one trial was removed from the situation and discarded from the experiment. On this basis, 3 anoxic animals and one control animal were discarded.

Thirty days after the original learning of the large maze had been completed, these animals relearned the maze to the same criterion (one errorless trial).

In the 30-day interval between the learning and relearning of the multiple unit maze, half (50) of the animals from each group of 100 learned a brightness discrimination problem in a single unit T water maze. The animals were run on the single unit maze 11 days after they had reached the criterion on the 14 unit maze.

The length of the stem of the T maze was 42 in.; the length of each arm of the T was 18 in. The ends of the arms of the T were extended forward 6 in. parallel with the stem, making it impossible to see the escape ladder from the choice point. The water was maintained at a depth of 9 in. The temperature of the water was 5-6 degrees F. below the temperature of the room.

At the choice point of the T, a septum was placed separating the arms. A rod passed through this septum and fastened to the sides of the stem was used to hang the black and white gates. The positions of the gates were varied randomly. The escape ladder was placed on the side of the "correct" gate. The gates were free swinging, permitting the animal to return to the choice point after making a "bad" choice.

Each animal was given 80 trials at this task extending over an 8 day period, viz., 8 blocks of 10 trials.

An additional 30 anoxic and 30 control animals learned the discrimination problem without previous experience on the multiple unit T maze. These performances (scored by trials and errors, only) were used to differentiate the groups on the basis of transfer of training from the maze learning to the brightness learning.

Most animals completed the learning and the relearning tasks by the 100th day of age. All animals were weighed at this time.

The schedule of testing with the various groups is shown in Table 14.

Results

The results are presented with reference to the following questions concerning the effects of prenatal exposure to oxygen deficiency upon learning and retention at maturity: (*a*) ability to learn and to retain the solution of a difficult maze problem, (*b*) relationship between the effects of prenatal anoxia on learning and the interval between the day of deprivation and the day of birth, (*c*) ability to learn a simple brightness discrimination problem, (*d*) transfer of training from one learning task to a second, and (*e*) inhibition (or facilitation) of interpolated learning upon relearning of the original task. Because of the imminent possibility of interaction effects, the analysis of the data also concerned questions of certain "side" effects of prenatal anoxia, that is, the effect on body weight, sex ratio, and speed of performance, specifically, speed in swimming the length of a straight tank.

TABLE 14

Program of Conditions of Learning for Each Group

Group	No.	Learning	Conditions of 30-Day Interval	Relearning
Control	100	Mult. **T** Maze	(50: rest) (50: learn B–W disc.)	Mult. **T** Maze
Anoxic	100	Mult. **T** Maze	(50: rest) (50: learn B–W disc.)	Mult. **T** Maze
Control	30	B–W discrimination		
Anoxic	30	B–W discrimination		

The effect of prenatal anoxia upon performance of learning and retention on difficult maze problem. The anoxic animals made 37% more errors in learning this maze to criterion (one errorless trial) than did the control group. The mean error score for the anoxic group was 226.33 and for the control, 164.61. The mean difference (61.72) was significant beyond the .01 level.

There were additional differences between the groups as related to the sex of the members. In superiority of performance (fewest errors), the control males were best; the control females, anoxic males, and the anoxic females followed in that order. The differences between the control males and the control females and between the control males and the anoxic males were significant beyond the .01 level. The difference between the control females and the anoxic females was significant beyond the .05 level (one-tailed test; the order of difference was as expected).

The anoxic animals required 30% more trials to reach the criterion of mastery than did the controls. The mean trial score of the anoxic groups was 25.23, whereas the corresponding score for the control group was 19.35; the difference between these means was significant beyond the .01 level.

As with the error scores, differences in trial scores also existed between the groups as related to the sex of the members. The nature of the differences was about the same: the control males were superior (required the fewest number of trials); the control females, the anoxic males, and the anoxic females followed in that order. The differences between the two groups of males and between the two control groups (male and female) were statistically significant (.01 level).

TABLE 15

Analysis of Error Data on 14 Unit Maze

Group	No.	Group Mean	σ_{Md}	Sex	t	p
Anoxic	100	226.33	15.37	Both	4.02	<.01
Control	100	164.61		Both		
Anoxic	39	241.85	24.86	F	1.89	<.05
Control	48	194.85		F		
Anoxic	61	216.41	18.82	M	4.24	<.01
Control	52	139.69		M		
Anoxic	61	216.41	25.17	M	1.01	>.05
Anoxic	39	241.85		F		
Control	52	136.69	18.40	M	3.15	<.01
Control	48	194.85		F		

TABLE 16

Analysis of Trial Data on 14 Unit Maze

Group	No.	Group Mean	σMd	Sex	t	p
Anoxic	100	25.23	1.82	Both	3.23	$<.01$
Control	100	19.35		Both		
Anoxic	39	27.05	2.81	F	1.59	$>.05$
Control	48	22.58		F		
Anoxic	61	24.07	2.31	M	3.33	$<.01$
Control	52	16.37		M		
Anoxic	61	24.07	2.98	M	0.90	$>.05$
Anoxic	39	27.05		F		
Control	52	16.37	2.30	M	2.70	$<.01$
Control	48	22.58		F		

The analysis of the time scores (time per trial) presents the same picture of inferiority of performance of the experimental animals. The anoxic group averaged 1.26 min. per trial as compared to the 1.16 min. averaged by the control group. The difference between these averages was statistically significant (beyond .01 level). Once again, further differences were noted when the two groups were subdivided according to the sex of the members. The control males were significantly faster than the control females (.05 level) and the anoxic males (.01 level).

As determined by performance on the multiple T maze, the differences between the groups were consistent for all measures used: for errors and trials to reach criterion and for average time per trial. In each instance, the control males were superior to the control females, anoxic males, and anoxic females. In each instance, differences between the control and experimental animals were attributable, largely, to differences between the two populations of male animals.

Retention of the solution of this multiple unit maze after an interval of 30 days following original mastery differentiated, significantly, the control and the experimental groups.

For this comparison of behavior, only half (50) of each of the two groups were used. (The remaining 50 of each group were used for comparison of the effects of interpolated learning.) The design and results are shown in Tables 17 and 18.

In all cases the raw relearning scores of the control group were superior to the scores of the anoxic group. The latter made 21% more errors and took 14% more trials in relearning the maze than did the former, the control group. Only the differences in trial scores were significant statistically.

These results might be a function of the greater amount of practice for the anoxic group in the original learning problem. The

TABLE 17

Comparative Retention Data for Control and Anoxic Animals under Normal ("Rest") Conditions for an Interval of 30 Days

(Error Scores)

Group	No.	Conditions during Interval	Group Mean	σMd	t	p
Anoxic	50	Rest	30.26	6.06	1.11	$>.05$
Control	50	Rest	23.52			

TABLE 18

COMPARATIVE RETENTION DATA FOR CONTROL AND ANOXIC ANIMALS UNDER NORMAL ("REST") CONDITIONS FOR AN INTERVAL OF 30 DAYS

(Trial Scores)

Group	No.	Conditions during Interval	Group Mean	σMd	t	p
Anoxic	50	Rest	6.50	0.88	1.77	<.05
Control	50	Rest	4.94			

anoxic animals made 46% more errors and required 34% more trials in the original learning situation. In any case, savings scores for the two groups were almost identical (median saving score in trials for control group, .80; median for anoxic group, .78).

Relationship between the effects of prenatal anoxia on learning and the interval between the day of deprivation and the day of birth. The data indicate significant differences in the ability to learn the maze as a function of the developmental stage at the time of oxygen deprivation. A summary of this analysis (analysis of variance) is given in Table 19.

A further review of the data revealed, however, that the bulk of the differences noted between the various experimental subgroups and between the total experimental group and the control group was contributed by a single experimental subgroup: the subgroup born 10 days after exposure (exposed, possibly, on the 13th fetal day). This subgroup differed significantly from each of the other experimental subgroups and from the control group. No differences existed among the other experimental subgroups. In a further comparison between the control and the anoxic animals, with the subgroup born 10 days after exposure deleted, significant differences remained although the level of significance was reduced (from .01 to .05).[10]

Ability to learn a simple brightness discrimination problem. Two groups of 30 animals, experimental and control, were compared to determine the effect of prenatal anoxia upon simple black–white discrimination learning. Neither group had been used previously on any learning problem.

The control group made more correct responses in the block of 80 trials than did the experimental. The mean number for

TABLE 19

ANALYSIS OF VARIANCE OF ANOXIC GROUPS WITH RESPECT TO THE INTERVAL BETWEEN DAY OF EXPOSURE AND DAY OF BIRTH

Age Groups	Measure	df	Sum of Squares	Variance	F	p
Between	Initial Learn.	8	280125.27	35015.66	2.69	<.01
Within	Errors	91	1183842.84	13009.62		
Between	Initial Learn.	8	3654.52	456.82	2.67	<.02
Within	Trials	91	15557.19	170.96		
Between	Relearning	8	2613.97	326.74	0.53	>.05
Within	Errors	91	62179.87			
Between	Relearning	8	165.84	20.73	1.27	>.05
Within	Trials	91	1506.91	16.38		

---

[10] Only the initial learning performance on the multiple maze could be evaluated in this fashion. Elsewhere the data were too meager to permit a subdivision according to fetal stage at time of deprivation.

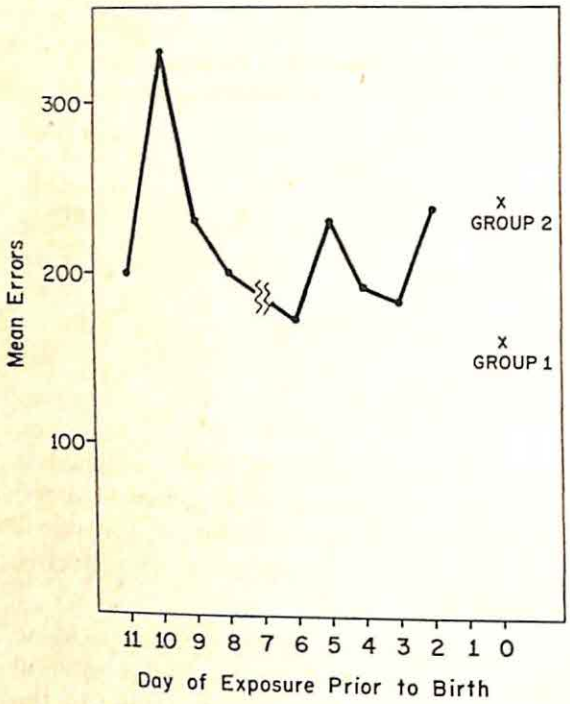


FIG. 11. Relationship between day of exposure to oxygen deficiency prior to birth and performance on the multiple T maze. For comparative purposes, the two points on the right, Group 1 and Group 2, are presented from the first study reported in this monograph.

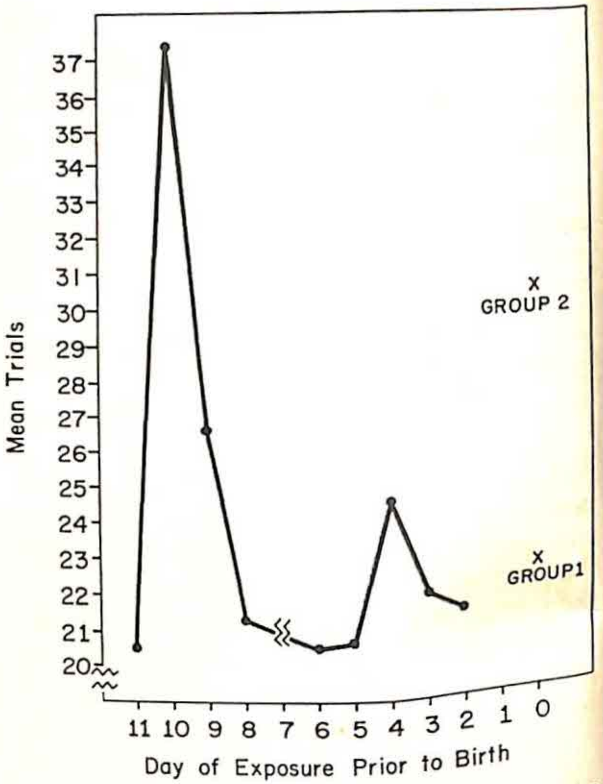


FIG. 12. Relationship between day of exposure to oxygen deficiency prior to birth and performances on multiple T maze. For comparative purposes, the two points on the right, Group 1 and Group 2, are presented from the first study reported in this monograph.

the control group was 50.93 and for the anoxic group, 47.17. These do not differ significantly, however ($p < .05$).

Transfer of training from one learning task to a second. Eleven days after they had reached the criterion on the multiple unit maze, 50 control and 50 anoxic animals were given 80 trials on the black–white discrimination task in the single unit water maze. The animals in both the anoxic and control groups, which were to be used on the discrimination problem, were selected prior to the beginning of the original learning on the complex maze. The selection was regarded as random with reference to the learning ability of the animals and to the subgroups related to the developmental stages at the time of deprivation.[11]

TABLE 20

ANALYSIS OF DATA ON SIMPLE DISCRIMINATION TASK: INITIAL LEARNING

Group	No.	Group Mean (CR)	σ*Md*	*t*	*p*
Anoxic Control	30 30	47.17 50.93	2.05	1.83	<.05

The results of their performances on this second learning problem, measured in terms of correct responses in the block of 80 trials, are given in Table 21. The control animals were superior to the experimental in both situations, that is, in situations with and without the benefit of previous learning experience (see Figure 13).

On the black–white discrimination problems, after maze experience, the control and the anoxic groups made 51.30 and 45.18 correct responses, respectively. This differ-

---

[11] As it turned out, there were differences in the learning on the initial task and in the representation of the several experimental subgroups.

TABLE 21

Analysis of Data on Simple Discrimination Task: Interpolated Problem

Group	No.	Group Mean (CR)	Sex	σMd	t	p
Anoxic	50	45.18	Both	1.65	3.71	<.01
Control	50	51.30	Both			
Anoxic	22	43.56	F	2.08	3.21	<.01
Control	25	50.14	F			
Anoxic	28	46.80	M	2.50	2.56	<.01
Control	25	53.21	M			

ence in discrimination learning, as before, favors the control group. This difference was significant at the .01 level. It appears that the ability to utilize the experience on the complex maze in learning of a discrimination problem, as well as the ability to master the complex maze, initially, was inferior in the anoxic animals.

Effects of interpolated learning upon relearning the original task. The conditions of interpolated learning were included in the experimental design in order to determine the relative susceptibility of the anoxic and the control animals to reproductive interference.

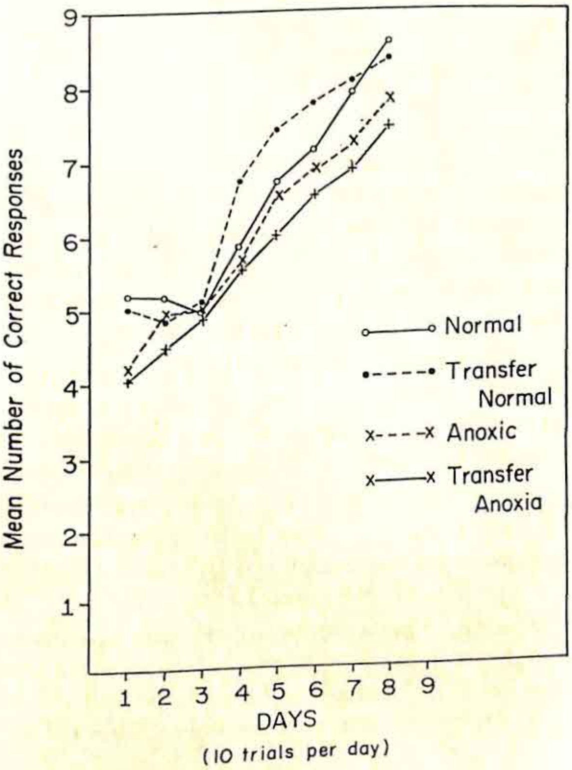


Fig. 13. Role of transfer of training among the control and experimental groups in the learning of the simple discrimination tasks.

The brightness discrimination problem, the interpolated task in the 30-day interval between the learning and relearning of the multiple unit maze, did not inhibit the relearning performances; instead, both the anoxic and the control groups demonstrated facilitation on the second exposure to the maze. The differences in relearning scores between the anoxic "rest" group and the anoxic "interpolated" group was significant beyond the .05 level. However, a comparable difference (.05 level) existed between the two anoxic groups on the original learning of the multiple unit maze. This existed despite a random selection of animals prior to any learning experience whatsoever. This latter difference tends to disclaim the reality of any difference between the two groups as related to interpolated experience.

Facilitation instead of inhibition from the interpolated learning may be explicable on the basis of a number of factors: the relative dissimilarity between the maze and the discrimination tasks in the specific stimulus properties critical to learning and upon which interference might depend; the relatively high similarity between the two learning situations in terms of the general conditions of practice and performance involved, on which facilitation might depend; and the mere occurrence of daily exercise, handling, or performance, even apart from learning, in a similar motivating situation for almost

TABLE 22

Effect of Interpolated Learning on Relearning

(Errors)

Group	No.	Conditions during Interval	Group Means	σ*Md*	*t*	*p*
Control	50	B–W discrimination	17.44	4.98	1.22	$>.05$
Control	50	Rest	23.52			
Anoxic	50	B–W discrimination	19.66	5.03	2.11	$<.05$
Anoxic	50	Rest	30.26			

$\frac{1}{3}$ of the 30-day interval which was not available to the "rest" groups.

"Side" Effects

Analysis of weight data. The mean weight of the anoxic group as a whole differs from the mean weight of the control group as a whole only on the basis of chance; no significant weight differences existed either at 60 or 100 days of age. A breakdown of the groups according to sex, however, revealed the significantly lighter weight of the anoxic females, as compared to the control females, at both dates. The anoxic and control males did not differ at any point. (See Table 24.)

Sex ratio. The subdivision of groups by sex was as follows:

	Multiple T Maze		Brightness Discrimination		Total	
	Male	*Female*	*Male*	*Female*	*Male*	*Female*
Control	52	48	15	15	67	63
Anoxic	61	39	20	10	81	49

Although in most instances there were more males than females, with the superiority in frequency magnified for the experimental groups, no significant differences existed.

Speed of swimming in the straight tank. From the absence of any differences in swimming speed, one may conclude that differences in learning performances of the anoxic and control groups cannot be attributed to this factor. The mean of the control group was 6.43 sec. and, of the anoxic group, 6.74 sec. per trial for the block of 10 trials. (The critical ratio on these data of .92 was insignificant.)

Even though the anoxic females were smaller than their control counterparts, as indicated by body weight, they were not materially affected in this aspect of locomotion.

Discussion

The primary result of this study is that animals exposed to severe oxygen deprivation for a short period of time during their fetal life are impaired at maturity in their ability to learn and to retain (after a 30-day interval) a complicated maze pattern. The

TABLE 23

Effect of Interpolated Learning on Relearning

(Trials)

Group	No.	Conditions during Interval	Group Means	σ*Md*	*t*	*p*
Control	50	B–W discrimination	4.22	0.73	0.99	$<.05$
Control	50	Rest	4.94			
Anoxic	50	B–W discrimination	4.00	0.79	3.16	$>.01$
Anoxic	50	Rest	6.50			

anoxic animals were inferior to the control animals in all measures of ability to learn and to retain the maze and to learn a discrimination problem, although in some instances this inferiority was not a statistically significant one.

The results of the initial learning on the 14 unit water maze were in the same direction as those results reported by Hurder and Sanders (1953) and Meier and Bunch (1950). In each of those studies, however, the deprivation occurred shortly after birth. In general, the results of this experiment indicate greater differences between the anoxic and control groups than in the other studies mentioned. It is impossible at this time to state the basis of the differences in magnitude between this and the other studies; it could be the prenatal (especially deprivation at a particular developmental stage) versus neonatal conditions; it could be a difference in the severity of the deprivation, per se; and so on.

Meier and Bunch reported that anoxic animals do not perform as well, according to some measures, as control animals on a single T water maze with respect to transfer of training in learning a position habit. It can be added from this study that animals deprived of oxygen prenatally: (*a*) do not perform as well as control animals on a single T water maze when ability to discriminate brightness is being measured and (*b*) show less ability to profit from previous learning of a maze problem.

The results on the relearning of the maze were not consistent with those reported by Meier and Bunch (1950), Windle and Becker (1943), and others reported in this monograph. Meier and Bunch reported that learning was relatively unaffected by neonatal deprivation, but that retention following 2 or 14 days was markedly altered. This study found that original learning is much more severely impaired than is retention. Retention here was measured by relearning (Meier & Bunch used both relearning and reversal of habit), which was started 30 days after the completion of the original learning. It is possible that the presence of significant differences may be a function of the length of the interval over which retention was measured.

A factor of unknown importance was the difference between the groups when subdivided by sex. Although not significant, a difference did exist in the frequency distribution of females and males between the control and experimental groups. In every instance where a comparison was possible, the males considerably outnumbered the females in the anoxic groups, but were almost equal in number in the control groups. Along with this, in each instance where significant differences were found between the experimental and control groups, the

TABLE 24

Analysis of Weight Data

Group	No.	Age (days)	Sex	Group Means (grams)	σMd	t	p
Anoxic	74	60	Both	167.58	5.36	1.25	$>.05$
Control	83	60	Both	174.27			
Anoxic	36	60	F	137.94	3.45	4.03	$<.01$
Control	42	60	F	151.83			
Anoxic	38	60	M	195.66	5.70	0.28	$>.05$
Control	41	60	M	197.24			
Anoxic	100	100	Both	245.99	6.78	0.02	$>.05$
Control	100	100	Both	246.11			
Anoxic	39	100	F	195.28	4.98	2.94	$<.01$
Control	48	100	F	209.93			
Anoxic	61	100	M	278.60	6.41	0.10	$>.05$
Control	52	100	M	279.21			

differences were largely explicable by the differences between the two populations (control and experimental) of males. This, however, was in spite of the consistently inferior performances of the anoxic females as compared to the other groups.

Although the anoxic females were significantly lighter than the control females (the only significant weight difference), the measures which would be expected to show consequential differences in behavior failed to do so. No differences were found in the swimming speeds in the straight tank. Differences in physical dimensions seemingly played an insignificant part in the explanation of differences in behavior.

In the light of these facts, there is a possibility that the learning data presented in this study were more favorable to the anoxic group than the differences discussed above would indicate. Possibly, one group, the females, was selected *against*. This group, as suggested by the nature of the behaviors found here, would have been even more drastically affected had a comparable ratio of females to males been attained for the anoxic group as was attained for the control group. Such an equitable distribution by sex, in all likelihood, would have made differences between the anoxic and control groups even greater.

The differences between the several experimental subgroups (subdivided by fetal age at the time of deprivation) are difficult to evaluate at this time. If one could make the necessary adjustments for species, developmental rates, severity of deprivation, and the time and nature of measurement, these differences might be found to be of the same basis as those described by Meier and Menzel (1955) in chicks. In their study, oxygen deprivation (LD_{50}) at the end of 8 days of incubation (21 days is "full term") was particularly effective in producing behavioral dysfunctions. The concepts offered by them in explanation of the results dealt with "critical period." In this usage, the concepts had been generalized from the critical periods in the growth of tissues and organs to critical periods in the growth of mature behavior patterns.

Several incidental, unquantified observations should also be added to the description of the experimental animals in this study. For the most part, none of the gross physical abnormalities reported by Ingalls (1947a, 1947b) in his experiments with mice were found with the rats in this study. It was noted, however, that 5 anoxic animals (of the 180 born) were blind and, consequently, were not used in the behavior measures.

Certain behavioral "peculiarities" were observed in the experimental animals. They tended to be more "skittish" and fractious than the controls. In their mastery of the several learning measures used, they tended to be more variable (this was reflected in the large group variances). These characteristics were similar to those observed in the study on neonatal anoxia in cats, reported by Meier (1953). Plausibly, these behaviors can explain differences in learning ability, especially since some of the differences tended to disappear as the animal had more experience in learning situations. That is, a deficit in learning ability may not be operative in these studies so much as a factor of heightened emotionality.

Summary and Conclusions

This experiment was conducted to measure the effects of prenatal anoxia on maze learning ability in rats. Forty pregnant rats were exposed in a decompression chamber at an altitude of 30,000 ft. (6.21% equivalent oxygen) for a period of 2 hrs. The rate of ascent and descent was 3,000 ft. per min. A total of 130 offspring of the exposed females were used in this study as experimental *S*s, and their performance on two different maze problems was compared to the performance of 130 offspring of 40 control animals.

The maze problems consisted of a 14 unit modified Stone **T** water maze on which control and experimental animals were required to reach a criterion of one errorless trial, and a single unit **T** water maze on which number of correct responses in 80 learning trials was the measure obtained.

Data were also procured for: retention as measured by relearning after a 30-day interval; transfer of training from the 14 unit T maze to the black–white discrimination problem on the single T maze; and the effects of learning an interpolated problem (the black–white discrimination problem) on relearning the original problem.

In addition to the above data, other variables considered were: weight of control and experimental animals at 60 and 100 days of age, speed of swimming in the two groups as measured by a straight water tank, the number of days before birth the anoxic animals were exposed to oxygen deficiency, and possible physiological disturbances caused by anoxia.

The following are conclusions based on the data obtained:

1. Prenatal anoxia severely affects an animal's ability to learn a 14 unit T maze. The animals deprived of oxygen for a short period during fetal life perform at maturity in an inferior manner to a control group in terms of errors, trials, and time. The differences were found to be significant beyond the .01 level.

2. Retention as measured by trials to relearn the 14 unit water T maze was significantly different between the control and experimental groups.

3. Anoxic animals which made the poorest learning records were those whose period of oxygen deprivation preceded birth by approximately 10 days.

4. The learning on a simple black–white discrimination problem on a single T water maze significantly differentiated the learning abilities of 30 anoxic and 30 control animals.

5. Positive transfer from the original learning problem to the simple black–white discrimination problem was slightly but significantly greater in the controls than in the anoxic animals.

6. The black–white discrimination problem, when interpolated between original learning and relearning of the 14 unit T maze, led to reproductive facilitation in both groups. The anoxic animals appeared to profit more from the interpolated learning than did the control group.

7. Generally, observable physical deformities were at a minimum. The only recognized deformity was blindness in five anoxic animals.

8. Female anoxic animals weighed significantly less than did female control animals; anoxic males did not differ from control males, even though the anoxic males were primarily responsible for significant differences between control and experimental groups on measures of behavior.

9. Speed of swimming in a straight tank was not seriously affected by prenatal oxygen deprivation.

10. The data do indicate that impairment in learning ability is a function of prenatal exposure to oxygen deficiency.

Histological Analysis

Following the completion of the maze tests, a limited sample of the animals was sacrificed and the brains prepared for histological analysis. The selection of the animals was made on the basis of their performances on the multiple unit maze. This sample included the two best and the four poorest performers of the control group and the four best and the six poorest performers of the experimental group (without regard to fetal age at the time of oxygen deprivation). In all, 16 *S*s were evaluated in this fashion.

In the data obtained, no significant histological trend, as related to prenatal anoxia, was apparent. No qualitative differences (size and shape of cells, etc.) could be seen between the control and experimental animals. While there were considerable variations in thickness of the cortex of the cerebrum and in the relative cell counts in the same structure, these seemed to be due, largely, to variations within the individual animal and in no way correlated with performance on the multiple unit maze.

Should prenatal anoxia cause cell anomalies or death, these techniques on this selected sample did not show them—certainly, these techniques did not show them to be functionally related to the measured behavior of the *S*s.

(For a more extensive discussion of the techniques used, their rationale, and the results gained, see Appendix B.)

Part II

Two questions raised by the first half of this study made a replication of the experiment, in part, mandatory. (*a*) Since the dates of conception were not known, the fetal ages at the time of oxygen deprivation could only be roughly estimated. Moreover, these estimates, based on the number of days prior to birth the fetuses were deprived, seemed risky, certainly questionable. Can this form of insult, an exposure to decompression with the attendant lack of oxygen, accelerate or, even, decelerate, prenatal development and alter, thereby, the time of parturition? (*b*) The maximal effect of the deprivation was upon those animals born 10 days after the insult. The effect upon the young born 9 days or less or 11 days or more after the insult was markedly less obvious. Could this in reality be a "critical period" at which time the developing organism is uniquely sensitive to changes in available oxygen? Could there be other critical periods at earlier fetal ages? (Apparently fetal ages less than 11–12 days were not sampled in the first part of this study.) These matters formed the basis of a replication in which the fetal ages tested would represent the entire period of gestation.

Subjects

According to plan, 66 pregnant female rats were to be used in this part of the study. Each of the days of gestation, from 1–22, were to be represented by 3 of the pregnant rats. Unfortunately, this plan met with some difficulty; only 62 pregnant Wistar rats were received from a local dealer at the appointed time. This meant that several of the fetal age groups were not fully represented (see Table 26). (The appearance of a vaginal plug was accepted by the breeder as positive indication of mating behavior.) Further, 2 of the females of the group which had been bred 22 days prior to their receipt in the laboratory dropped their young within minutes of the scheduled time of the oxygen deprivation. Consequently, only 21, rather than 22, intervals in the fetal development were examined in this study.

Apparatus and Procedure

Deprivation. Forty-one of the pregnant females, 2 from each of the 21 groups of 3, were placed in a large decompression chamber. (There were exceptions; these are indicated in Table 25. Where both the control and experimental groups could not be included in a fetal age group, because of small size, only the experimental group was represented. As a result of this, there were 41 experimental pregnant animals and 19 control.) The pressure within the chamber was reduced until a simulated altitude of 30,000 ft. (8.88 in. Hg.; 6.21% equivalent oxygen) was reached. This simulated altitude was reached in 15 min. The animals were maintained under these conditions for 120 min., after which time the pressure was returned to normal. This, too, required 15 min. All in all they were under some degree of decompression and oxygen deprivation for 150 min.

The females were placed in individual cages and the respective dates of delivery of their litters were recorded.

Of the 254 young born to the experimental females and the 124 born to the controls, 108 were selected for testing purposes. Since it was recognized that litter differences might be a major confounding factor in this study (with the small number of litters at each fetal age), the experimental offspring were divided into five groups. These groups, with their numerical size, were:

Group I: Deprived 2–5 days prior to birth: 14 *S*s.

Group II: Deprived 6–9 days prior to birth: 14 *S*s.

Group III: Deprived 10–13 days prior to birth: 19 *S*s.

Group IV: Deprived 14–17 days prior to birth: 20 *S*s.

Group V: Deprived 17–21 days prior to birth: 16 *S*s.

Control Group: 25 *S*s.

(Originally, it was planned that there be but 14 *S*s in each of the 5 experimental groups and 25 in the control group. As the running of the *S*s on the maze was more rapid than anticipated, the size of the last three experimental groups was increased.) The *S*s for these groups were selected with the view of representing as many individual litters as possible. No consideration was given to the sex ratios of the several groups. In all, there were 83 *S*s from the asphyxiated mothers and 25 from the control used for testing purposes.

Behavior measures. The 108 animals were permitted to mature under normal laboratory conditions until the 64th–67th days at which time they were pretrained on the straight swimway used in the first half of this study. After 10 trials on the swimway (completed within a single session) the animals were run on the multiple unit water maze (also used in the first portion of the study) until they reached the criterion of one errorless performance. On this apparatus the *S*s were given 3 trials each day. Besides trials to criterion, time and error scores were recorded.

Results

Vital statistics. Of the 41 pregnant animals exposed to the deprivation conditions, 3 died during or immediately following exposure. Two more died within 5 days of the exposure. There seemed to be no pattern in these deaths; deaths occurred within the entire range of gestation ages (see Table 25). (No control females died within this period.)

Only 44 litters were delivered; 30 to the experimental group and 14 to the control. Again, no pattern can be found in these failures; in fact, the notion that some of the experimental animals could have aborted does not seem to be tenable, since the number of failures for the two groups are almost the same (5, control; 6, experimental).

Although there was a difference in the average length of the gestation periods in favor of the control group (0.42 days), this was not found to be significant. (See Table 26.) More than likely, this difference is considerably less than the errors inherent in the setting of the date of mating and the date of delivery. A cursory inspection of Table 28 will reveal that only one delivery seemed to have an unusually short gestation period and only two longer (considerably) than usual. Other than these, the control and the experimental groups were very much alike.

Again, although there were differences in litter size at birth and at 65 days of age in favor of the control group, statistical significance of these differences was lacking.

TABLE 25

VITAL STATISTICS: I

(Breeding and delivery data on the control (C) and experimental (X) females)

Mating Date	No. in Group	No. of Litters		No. of Deaths		Failures to Drop	
		C	X	C	X	C	X
8–1	1	1					
8–2	3	1	2				
8–3	3	1	2				
8–4	3	1	1		1		
8–5	1		1				
8–6	2		1		1		
8–7	3		2			1	
8–8	3	1	2				
8–9	3	1	2				
8–10	3	1	2				
8–11	3		1			1	1
8–12	3	1			1*		1
8–13	2		2				1
8–14	3	1	1		1*		
8–15	3	1	2				
8–16	3	1	2				
8–17	3	1	2				
8–18	3	1	2				
8–19	3		2			1	
8–20	3				1*	1	1
8–21	3		1			1	1
8–22	3	1					2
Totals	60	14	30		5	5	6

* Died during the oxygen deprivation. The others died within 5 days of exposure.

TABLE 26

VITAL STATISTICS: II

Litter and Viability Data with Respect to Dates of Mating and Parturition

Birth Date	Breeding Date	Length of Gestation (days)	Group	No. Born	No. Dead at Birth	No. Dead at 65 Days	No. Alive at 65 Days	Dispensation
8–23	8–1	22	C	14			14	6 to C
8–25	8–2	23	X	9			9	5 to I
	8–2	23	C	9	1	1	7	
	8–3	22	X	5	1	1	3	1 to I
8–26	8–2	24	X	8	4	3	1	
	8–3	23	X	3		3		
	8–3	23	C	9		1	8	2 to C
8–28	8–4	24	X	7	3		4	2 to I
	8–4	24	C	12			12	4 to C
	8–5	23	X	6			6	3 to I
	8–6	22	X	6			6	3 to I
8–30	8–7	23	X	13	1	2	10	4 to II
	8–7	23	X	11		10	1	
	8–8	22	C	3		1	2	1 to C
	8–8	22	X	6		6		
	8–8	22	X	11	1	1	10	4 to II
8–31	8–9	22	X	10			10	6 to II
	8–9	22	X	11		8	3	6 to 2
9–1	8–9	23	C	9	1	1	7	2 to C
9–2	8–10	23	C	3	2	1		
9–3	8–10	24	X	12	1	2	9	6 to III
	8–11	23	X	10		7	3	
9–4	8–12	23	C	10		5	5	2 to C
	8–15	20	X	14	2	1	11	5 to III
	8–13	22	X	11		2	9	6 to III
9–5	8–13	23	X	7			7	2 to III
	8–14	22	C	10			10	3 to C
9–6	8–10	27	X	7	3		4	3 to IV
	8–14	23	X	11	5	3	3	3 to IV
	8–15	22	C	6		6		
9–7	8–15	23	X	11		2	9	5 to IV
9–8	8–17	22	C	8		6	2	
9–9	8–16	24	X	3		1	2	1 to IV
	8–16	24	X	13		1	12	8 to IV
	8–17	23	X	8	1	7		
	8–18	22	C	10	2	8		
9–10	8–16	25	C	9			9	3 to C
	8–17	24	X	3	3			
	8–18	23	X	9	1	1	7	2 to V
9–11	8–19	23	X	5			5	5 to V
9–12	8–19	24	X	4		4		
9–13	8–18	26	X	13	1		12	5 to V
	8–21	23	X	7	2	1	4	4 to V
	8–22	22	C	9		2	7	2 to C
Number (control)				121	6	32	83	
Number (experimental)				254	28	66	160	
Mean (control)				8.64			7.54	
Mean (experimental)				8.47			6.40	

Any differences between the two groups in the matter of litter size was attributable to difference in the litters at the time of birth. Further, there was no trend in the size of the litter and the gestation age at the time of deprivation.

In summary, with the procedures used in which the breeding and deliveries were checked regularly once every 24 hrs., no differences could be detected between the control and the experimental groups of animals. Possibly, more frequent inspections could have shown differences between the two groups.

Behavioral measures: Pretraining and maze performance. The performances of the animals were analyzed first by treatment, that is, by age at the time of deprivation, then by sex. As on most of the criterion measures the data were nonnormal; these analyses were completed with the use of the Kruskal-Wallis (Siegel, 1956, pp. 184-193) one-way analysis of variance. This particular analysis utilizes the data by ranks rather than by the actual raw data. (In these instances there was little loss of statistical power as compared to the usual F tests. Were the assumptions of F met, the asymptotic efficiency of the Kruskal-Wallis would be 95.5.)

The median time scores for the 10 trials on the straight swimway were not statistically significant whether analysed by age at deprivation or by sex. The median scores were:

	C	I	II	III	IV	V	Totals
Males	48.4	51.4	63.9	50.0	51.3	63.5	54.3
Females	52.4	53.8	72.0	51.3	53.4	49.8	54.7
Totals	50.3	52.7	67.4	50.6	52.6	58.3	

The variance H for treatments was 3.47; for sex, 0.00. (For significance at the .05 level, H would have to equal or exceed 12.83 for treatments and 5.02, for sex.)

On the maze itself, seven different criterion measures were evaluated. These were: (*a*) trials to criterion, (*b*) errors to criterion, (*c*) errors per trial, (*d*) errors on the first three trials, (*e*) time to criterion, (*f*) time per trial, and (*g*) time on the first three trials.

Trials to criterion: no differences were found between groups on the number of trials required to reach the criterion of mastery (one errorless performance). The median performances were:

	C	I	II	III	IV	V	Totals
Males	55.9	44.6	49.6	47.3	48.7	60.2	51.8
Females	51.4	67.1	69.0	49.9	48.0	73.3	57.5
Totals	53.7	57.4	58.0	48.5	48.3	65.1	

Both the H for treatment, 3.60, and the H for sex, 0.92, were clearly below the required values (12.83 and 5.02, respectively) for significance at the .05 level.

Errors to criterion: although the various treatment groups could not be differentiated on this measure of performance, the males and females were found to be significantly different ($p < .05$) on the number of errors made in reaching mastery. The males were vastly superior to the females in all treatment groups.

	C	I	II	III	IV	V	Totals
Males	58.2	47.8	57.5	47.1	30.9	42.2	47.5
Females	59.8	62.1	86.2	59.2	50.4	68.7	62.4
Totals	59.0	56.0	69.8	52.8	40.6	52.1	

H for treatment was 7.95; for sex, 6.09. (For significance at the .05 level this latter value must equal or exceed 5.02.)

Errors per trial: again, no differences were found between groups due to treatment on the basis of the ratio of total errors to total trials to reach mastery for each *S*. Nor were differences found in performance due to sex.

	C	I	II	III	IV	V	Totals
Males	56.2	63.4	67.9	56.8	31.7	36.4	50.9
Females	63.9	48.6	65.1	67.7	61.8	35.2	58.5
Totals	59.9	54.9	66.7	62.0	46.7	35.9	

H for treatments was 10.8; for sex, 1.57.

Errors on the first three trials: since one of the *S*s mastered the maze in three trials, the data for all were analyzed through this point. By this we hoped to get a common base line for all *S*s unbiased by "obvious relationships" as would be expected to occur in the derived measures of behavior, e.g., errors per trial, time per trial.

The groups did differ significantly ($p < .05$) by treatments on the number of errors made in the first three trials. No differences were shown in behavior attributable to sex.

	C	I	II	III	IV	V	Totals
Males	51.8	71.3	69.3	47.0	37.5	42.9	51.4
Females	48.6	67.0	82.2	63.9	49.9	47.2	58.0
Totals	49.8	68.8	74.8	55.0	43.7	44.5	

H for treatments was 13.41; for sex, 1.19. (For significance at the .05 level, *H* must equal or exceed 12.83.)

Time to criterion: again, the groups differed significantly ($p < .05$) by treatments, now in the total time required for the mastery of the task. Again, no differences in performance were found attributable to sex differences.

	C	I	II	III	IV	V	Totals
Males	65.5	57.8	53.1	34.3	37.3	50.6	49.9
Females	56.3	74.5	81.2	50.9	39.3	72.0	59.6
Totals	61.1	67.3	65.1	42.2	38.3	58.6	

H for treatments was 13.65; for sex, 2.60.

Time per trial: as might be expected from the significant differences in total time and the lack of significance in total trials (although the two trends were the same), a derived measure based upon these two also pointed to significant differences ($p < .01$) between the groups by treatment. No differences in performance were related to differences in sex.

	C	I	II	III	IV	V	Totals
Males	67.4	71.8	62.8	58.7	43.1	45.2	49.5
Females	64.3	74.5	77.7	35.7	29.2	36.6	60.0
Totals	65.9	73.4	69.1	46.6	36.1	39.8	

H for treatments was 23.06; for sex, 3.03. (For significance at the .01 level, *H* for treatments must equal or exceed 16.75.)

Time on the first three trials: the groups, as classified by dates of delivery after the period of oxygen deprivation, differed significantly ($p < .02$) on the times to complete the first three trials. Once more, no differences in time scores were attributable to differences in sex.

	C	I	II	III	IV	V	Totals
Males	65.7	72.1	67.4	35.6	46.0	49.0	54.9
Females	51.4	78.3	68.6	49.4	42.1	39.4	54.0
Totals	58.8	75.6	67.9	42.1	44.0	45.4	

H for treatments was 16.00; for sex, 0.02. (For significance at the .02 level, *H* for treatments must equal or exceed 15.09.)

Discussion

On the face of things, the data on duration of gestation justifies an extrapolation of the gestation age at the time of deprivation from the number of days which this deprivation preceded birth, as might have been done in the first part of the study. However, there are several qualifications that must be considered before concluding that this bout of oxygen deprivation has no effect on the duration of gestation. The first deals with the procedures used in this study for the determination of the data of breeding and delivery. Possibly, a closer check could have been made on the time of mating. From the breeder, we learn that his procedures entailed a maximum error of 12 hrs. Certainly, a more frequent check could have been made in the designation of the date of delivery. Routine 12 hr., rather than 24 hr., observations would have been feasible. In all, more frequent observations could have materially reduced the procedural variance which in this instance was so much larger than the group differences.

Second, work in the Vanderbilt Laboratory has shown a regular variation in the date (and hour) of hatching of chicks which had been exposed to an acute oxygen deficiency during incubation. We have found that a median lethal deficiency at the end of the 8th day of incubation (full incubation: 21 days) will delay hatching by an average of 12 hrs., whereas deprivation at the end of the 19th day will accelerate hatching by 6 hrs. (Again, chicks are uniquely suited to this type of research; here, because the determination of date of "birth" is both easier and more meaningful for the individual *S*.)

Although significant differences were detected among treatment groups only on the last four measures, their pattern of performance was consistent on all seven. The poorest performances were observed in animals born 6–9 days following oxygen deprivation (gestation age: 13–16 days). The best performances were observed in those animals born 14–17 days following deprivation (gestation age: 5–8 days). In broad outline, these data in part 2 are consistent with those of part 1.

In some respects, however, the notion of critical periods was not materially strengthened. A cursory observation of the group trends, both on separate and pooled measures, reveals an increased susceptibility to impairment by this insult as the animals mature to parturition. Nevertheless, it should be pointed out that the performances of the animals deprived very *early* in gestation were as different from the controls as those of the animals deprived very *late*. On most measures, the control group stood in between these two age groups.

Conceivably, some factor could be affected in early gestation which would tend to "improve" the *S*s' behaviors, as measured here. (At least, the time scores were shortened, the error scores were kept to a minimum, etc.) Likewise, some factor could be affected in middle and late gestation which would tend to "downgrade" the *S*s' behaviors, again, at least as measured here. Given the two separate functions for these factors, one could justifiably speak of two critical periods centering about the two phases of the animals' prenatal development.

Summary and Conclusions

This replication was made to check on the effect of oxygen deprivation on the length of gestation and the possibility of "critical periods" earlier in gestation than noted in the earlier part of this study.

Forty-one pregnant Wistar rats were exposed to a simulated altitude of 30,000 ft. for 120 min. These animals represented, in groups of two (with irregularities, sometimes less), 21 stages in the fetal development of the rat. Following exposure, the animals were placed in individual cages and the dates of delivery recorded.

The progeny of 19 pregnant animals of comparable gestational development were the controls.

From the data collected, no differences would be ascribed to the experimental and control groups with reference to duration of gestation, size of litter at birth, and size of litter at time of running on the behavior measures.

At the age of 64–67 days the animals were run 10 trials on a straight swimway following which they were run on a 14 unit water maze to the criterion of one errorless trial. On the basis of the performances on these apparatus we can conclude:

1. There were no differences in median time scores for the 10 trials of the straight swimway. This might be construed to mean that there were no differences in activity levels between the experimental and control groups.

2. No differences were noted between the groups on the number of trials or errors to reach the criterion of one errorless performance. Further, no differences were found in the number of errors per trial made by the two groups of *S*s.

3. Significant differences were found in the time measures of behavior, the time to criterion ($p < .05$), and the time per trial ($p < .01$).

4. As one *S* (control) mastered the maze in three trials, analyses of the behaviors of all of the *S*s were made through this point. There were significant differences on both the number of errors made in these three trials ($p < .05$) and in the time required for completion of these trials ($p < .02$).

5. Consistently, Group II made the poorest performances. This group was comprised of animals born 6–9 days following the oxygen deprivation (gestation age: 13–16 days). This tends to confirm the findings of part 1 of this study.

6. Consistently, Group IV made the best performances. This group was comprised of animals born 14–17 days following the deprivation. (Gestation age: 5–8 days.) Conceivably, the performances of this group and of Group II indicate the existence of two unique "critical periods." As measured in this study, an insult in one period may have a facilitative effect whereas an insult in the other may have a deleterious effect on maze learning.

SUMMARY AND CONCLUSIONS

The three experiments reported in the present study are primarily concerned with the effects of natal and prenatal anoxia upon subsequent behavior, and upon learning and retention at maturity in rats and in cats.

Results are reported on the effects of a single exposure to anoxia, at birth in experiment 1 and at different stages during the prenatal period in experiment 3, on the ability of white rats at maturity (*a*) to learn a difficult and reliable maze problem (14 unit multiple T pattern), (*b*) to utilize the previous learning of the maze problem in the later learning of a simple discrimination task (transfer of training), (*c*) to retain the maze habit over an interval of 30 days under normal conditions during the interval, (*d*) to retain the maze habit over an interval of 30 days under conditions designed to produce retroactive inhibition, (*e*) to extinguish a bar pressing response recently acquired, and (*f*) to reverse a simple position habit shortly after its mastery. In all, data are reported on approximately 474 white rats in the learning and retention studies. Approximately one half were control and one half were anoxic animals.

Behavior data are presented in experiment 2 on cats showing the effects of a single exposure to anoxia at birth upon their ability to learn, as measured in the following battery of tasks: Guthrie puzzle box, single unit T maze, double alternation problem, and a brightness discrimination problem. Also investigated, but without positive results, were the effects of natal anoxia upon adult EEG patterns in the cat and the influence of prenatal anoxia upon gestation period and upon structural brain damage in adulthood in the rat.

The experiments have shown the following results:

1. *Natal Anoxia in Rats*

a. Thirty minutes of oxygen deprivation (2.91% equivalent oxygen) following birth had no significant effect upon learning or retention in rats of a difficult maze problem nor upon the extinction of a bar pressing habit.

b. Sixty minutes of oxygen deprivation under similar conditions, on the other hand, produced significantly inferior learning and retention of the multiple unit maze, in terms both of number of errors made and number of trials required to reach the criterion of mastery.

c. Sixty minutes of the severe oxygen deprivation did not significantly affect the learning of a simple position response but it did have a deleterious effect upon the reversal of that habit.

2. *Natal Anoxia in Cats*

a. Rapidity of response in the puzzle box was significantly decreased.

b. Poorly adapted responses, such as "stereotypy," were significantly increased in certain discrimination tasks: learning a position response rather than the required discrimination.

c. A reduced ability to sustain a symbolic solution once the solution had been reached.

d. There was no indication that oxygen deprivation affected the ability to learn a brightness discrimination where there was close spatial relationship and temporal proximity of response between instrumental objects and reward.

3. *Prenatal Anoxia in Rats*

In Part I, in which 40 pregnant rats were exposed to 6.21% equivalent oxygen for two hrs. and 130 offspring were compared with 130 offspring of 40 control mothers, the effects of prenatal anoxia upon learning, retention, and transfer were as follows:

a. The animals deprived of oxygen during fetal life performed in a significantly inferior manner to control animals in terms of errors, trials, and time required to learn the multiple unit maze.

b. Retention of the maze problem, as measured by trials to relearn, was significantly impaired.

c. Learning the simple black–white discrimination problem was significantly poorer in the anoxic as compared with the control animals.

d. Slight but significantly less positive transfer was shown by the anoxic than by the control animals from the maze problem to the black–white discrimination problem.

e. Anoxic animals which made the poorest learning records were those whose period of oxygen deprivation preceded birth by approximately 10 days.

f. Retroactive inhibition in the retention of the maze problem was not produced by the interpolated discrimination learning, in either the anoxic or control animals.

g. Observable physical deformities were at a minimum; the female anoxic animals weighed significantly less than did female control animals.

Part II replicated Part I with the modification that the single exposure of pregnant females of 6.21% equivalent oxygen for a period of two hrs. was studied at 21 different stages in the gestation period. The comparison of the prenatal anoxic group with control animals showed the following results:

1. There were significant differences between anoxic and control animals indicating superiority of the latter in the following measures in learning the maze problem: the time to criterion, the time per trial, and number of errors made on the first three trials. There were no effects of anoxia in the number of trials or errors made in learning the problem to mastery.

2. The gestation age at which oxygen deprivation occurred was a significant factor with respect to maze learning ability. Animals born 6–9 days following deprivation (gestation age: 13–16 days) consistently made the poorest performance. Animals born 14–17 days following deprivation (gestation age: 5–8 days) consistently made the best performance.

3. Vital Statistics: of the 41 pregnant animals exposed to the deprivation conditions, three died during or immediately following exposure and two more died within five days of the exposure. Mortality did not seem to be related to a particular gestation age at time of deprivation. No control females died during this period.

4. No differences were found between anoxic and control groups regarding duration of gestation period.

REFERENCES

Apgar, Virginia, Girdany, B. R., McIntosh, R., & Taylor, H. C., Jr. Neonatal anoxia: I. A study of the relation of oxygenation at birth to intellectual development. *Pediatrics,* 1955, **15,** 653–662.

Armitage, S. G. The effects of barbiturates on the behavior of rat offspring as measured in learning and reasoning situations. *J. comp. physiol. Psychol.,* 1952, **45,** 146–152.

Becker, F. R., & Donnell, W. Learning behavior in guinea pigs subjected to asphyxia at birth. *J. comp. physiol. Psychol.,* 1952, **45,** 153–162.

Breslin, F. D. The effect of oxygen lack upon physical and neurological development in the chick. *Amer. J. Ment. Defic.,* 1956, **60,** 595–602.

Charles, Margaret S., & Fuller, J. L. Developmental study of the electroencephalogram of the dog. *Electroenceph. clin. Neurophysiol.,* 1956, **8,** 645–652.

Despert, J. L. Anxiety, phobia, and fears in young children. *Nerv. Child,* 1946, **5,** 5–23.

Edwards, A. L. *Experimental design in psychological research.* New York: Rinehart, 1950.

Fazekas, J. F., Alexander, F. A. D., & Himwich, H. E. Tolerance of the newborn to anoxia. *Amer. J. Physiol.,* 1941, **134,** 281-287.

Fisher, R. A. *Statistical methods for research workers.* London: Oliver & Boyd, 1946.

Gibbs, F. A., & Gibbs, Erna L. *Atlas of electroencephalography.* Vol. I. *Methodology and controls.* (2nd Ed.) Cambridge, Mass.: Addison-Wesley, 1950.

Glass, H. G., Snyder, F. F., & Webster, E. The rate of decline in resistance to anoxia in rabbits, dogs, and guinea pigs from the onset of viability to adult life. *Amer. J. Physiol.,* 1944, **140,** 609–615.

Gomez, L., & Pike, F. H. The histological changes in nerve cells due to total temporary anaemia of the central nervous system. *J. exp. Med.,* 1909, **11,** 257–265.

Grenell, R. G. Central nervous system resistance: I. The effects of temporary arrest of cerebral circulation for periods of two to ten minutes. *J. Neuropath. exp. Neurol.,* 1946, **5,** 131–156.

Guthrie, E. R., & Horton, G. P. *Cats in a puzzle box.* New York: Rinehart, 1946.

Hamilton, H. L. *Lillie's development of the chick.* New York: Holt, 1952.

Harlow, H. F. Primate learning. In C. P. Stone (Ed.), *Comparative psychology.* (3rd ed.) New York: Prentice-Hall, 1951. Pp. 183–238.

Himwich, H. E. *Brain metabolism and cerebral disorders.* Baltimore: Williams & Wilkins, 1951.

Himwich, H. E., & Fazekas, J. F. Comparative studies of the metabolism of infant and adult dogs. *Amer. J. Physiol.,* 1941, **132,** 454–459.

Hurder, W. P. Quantitative cortical changes produced by anoxia in rats. *J. comp. physiol. Psychol.,* 1952, **45,** 362–367.

Hurder, W. P., & Sanders, A. F. The effects of neonatal anoxia on the maze performance of adult rats. *J. comp. physiol. Psychol.,* 1953, **46,** 61–63.

Ingalls, T. H. Etiology of mongolism. *Amer. J. Dis. Child.,* 1947, **73,** 147–165.

Ingalls, T. H. Pathogenesis of mongolism. *Amer. J. Dis. Child.,* 1947, **73,** 279–292.

Jasper, J. H., Bridgman, C. S., & Carmichael, L. An ontogenetic study of cerebral electrical potentials in the guinea pig. *J. exp. Psychol.,* 1937, **21,** 63–71.

Johnson, P. O. *Statistical methods in research.* New York: Prentice-Hall, 1949.

Jones, L. V. Tests of hypotheses: One-sided vs. two-sided alternative. *Psychol. Bull.,* 1952, **49,** 43–46.

Kendall, M. G. *Rank correlation methods.* London: Charles Griffin, 1948.

Levinson, Billey. Effects of fetal irradiation on learning. *J. comp. physiol. Psychol.,* 1952, **45,** 140–145.

Libet, B., Fazekas, J. F., & Himwich, H. E. The electrical response of the kitten and adult cat brain to cerebral anemia and analeptics. *Amer. J. Physiol.,* 1941, **132,** 232–238.

Licklider, J. C. R., & Bunch, M. E. Effects of enforced wakefulness on growth and maze learning performance of white rats. *J. comp. physiol. Psychol.,* 1946, **39,** 339–350.

Little, W. J. On the influence of abnormal parturition, difficult labours, premature birth, and asphyxia neonatorum on the mental and physical condition of the child, especially in relation to deformities. *Trans. obstet. Soc. Lond.,* 1861, **3,** 293–344.

McNemar, Q. *Psychological statistics.* New York: Wiley, 1949.

Meier, G. W. Delayed effects of natal anoxia upon behavior and electroencephalographic activity. Unpublished doctoral dissertation, Washington University, St. Louis, 1953.

Meier, G. W. Prenatal anoxia in relation to behavioral phenomena in other animals. In W. F. Windle (Ed.), *Neurological and psychological deficits in asphyxia neonatorum.* Springfield, Ill.: Charles C. Thomas, in press.

Meier, G. W., & Bunch, M. E. The effects of natal anoxia upon learning and memory at maturity. *J. comp. physiol. Psychol.,* 1950, **43,** 436–441.

MEIER, G. W., & MENZEL, E. W. Prenatal oxygen deprivation and subsequent behavior dysfunctions. *Science,* 1955, **122,** 419–420.

MOSES, L. E. Non-parametric statistics for psychological research. *Psychol. Bull.,* 1952, **49,** 122–143.

PALMER, F. H. Comparative performance of cats and rats in double alternation problem of the temporal maze. Paper read at Midwest Psychological Association, Detroit, 1950.

PRESTON, MARY I. Late behavioral aspects found in cases of prenatal, natal, and postnatal anoxia. *J. Pediat.,* 1945, **26,** 353–366.

ROSENFELD, G. B., & BRADLEY, C. Childhood behavior sequelea of asphyxia in infancy, with special reference to pertussis and asphyxia neonatorum. *Pediatrics,* 1948, **2,** 74–84.

ROYCE, J. R. The factorial analysis of animal behavior. *Psychol. Bull.,* 1950, **47,** 235–259.

SCHEIDLER, C. H. The effects of prenatal anoxia on learning of white rats. Unpublished doctoral dissertation, Washington University, St. Louis, 1953.

SCHREIBER, F., & GATES, N. Cerebral injury in the newborn due to anoxia at birth. *J. Mich. State med. Ass.,* 1938, **37,** 145.

SEWARD, J. P. The sign of a symbol: A reply to Professor Allport. *Psychol. Rev.,* 1948, **55,** 277–296.

SIEGEL, S. *Nonparametric statistics.* New York: McGraw-Hill, 1956.

SNYDER, F. F. *Obstetric analgesia and anesthesia.* Philadelphia: Saunders, 1949.

VAN LIERE, E. J. *Anoxia: Its effect on the body.* Chicago: Univer. Chicago Press, 1942.

WINDLE, W. F., & BECKER, R. F. Asphyxia neonatorum. An experimental study in the guinea pig. *Amer. J. Obstet. Gynec.,* 1943, **45,** 183–200.

WINDLE, W. F., BECKER, R. F., & WEIL, A. Alterations in brain structure after asphyxiation at birth. An experimental study in the guinea pig. *J. Neuropath. exp. Neurol.,* 1944, **3,** 224–238.

YOUNGDAHL, PATRICIA L. The effects of natal anoxia on the learning ability of white rats. Unpublished master's thesis, Washington University, St. Louis, 1948.

(Received October 17, 1958)

APPENDIX A

ELECTROENCEPHALOGRAPHIC DATA

Even though the analysis of electroencephalographic frequencies and amplitudes for differences in the normal and natally anoxic groups of cats proved negative, certain regularities appeared in both groups as related to the development of electrical activity of the brain.

The frequencies of the EEGs through maturation typically increased until the third or fourth month when the adult level was reached. At birth the dominant frequency was less than 1 cycle per second (cps) with occasional "runs" of greater frequencies. At the end of the first month, 3–5 cps activity was typical; at the end of the second month, 5–7 cps; and at the end of the third or fourth month, 6–9 cps. These last frequencies were those found at the end of the sixth month and, again, at the final recording: 7–9 cps activity was representative of the frontal leads and 7–8 cps of the occipital.

This listing of "dominant" frequencies should not be construed to mean that these were the only frequencies seen at the particular recording. The spectra of frequencies were much the same for all ages, that is, both slow (1–2 cps) and fast (10–15 cps) appeared in the records at all ages. The frequencies listed as dominant are those which appeared most frequently in the record of a given *S* for a given age. This does not imply, further, that these frequencies constituted necessarily the majority of the record.

Dominance appeared in the occipital leads beginning with the third recording (end of the second month). Almost universally, the left occipital records had greater amplitudes than the right.

As indicated earlier (cf. *Apparatus and Procedure*) the electroencephalographic recordings at the time of deprivation demonstrated a pattern of activity correlated with the severity and duration of anoxemia. Prior to the deprivation the frequencies were typically less than 1 cps within frequent sequencies of 3–5 cps activity. During the earlier stages of deprivation, about the stage when the equivalent oxygen in the chamber was reduced by half, the frequencies increased and regular sequences of 3–5 cps activity prevailed. This was concurrent with increased motor activity of the *S*; the increased electrical activity appeared to be causally related to motor action. After this point, the frequency and amplitude diminished continuously until the deprivation bout was concluded. From many of the *S*s, no activity could be recorded at this time; that is, by this instrumentation both the frequency and the amplitude were reduced to "zero" (5 microvolts produced $\frac{1}{16}$ in. deflection). Occasionally, however, some of these periods of "no activity" were disturbed by high amplitude (high frequency bursts that were coincidental with gasping behavior), the last reflex to be abolished. Whether any relationship existed between the "absence" of electrical activity (rather than a mere diminution of frequency and amplitude) at this time and defective behavior later could not be discerned from the data collected in this study.

APPENDIX B

HISTOLOGICAL TECHNIQUE AND RESULTS

The preparation of the material for histological examination of permanent changes attributable to prenatal anoxia was as follows:

The animals were first anesthetized (40 mg. pentobarbital/kg. body wt., administered intraperitoneally) and then killed by perfusion with an isotonic saline solution followed by a 10% solution of formaldehyde.

The brains were removed and stored for at least two weeks in the 10% formaldehyde solution before preparation for sectioning was begun. Save for the cerebellum with the attached pons and medulla, the brain was cut transversely to the long axis, perpendicular to the surface of the cerebrum. For the purposes of this analysis, two consecutive sections (cut at 10 micra) out of each 16 were mounted and stained with thionin, a Nissl stain. However, only the sections through the cerebrum in the region of the thalamus were examined.

The rationale for this procedure was founded on certain facts of neuronal degeneration and on certain findings reported by other investigators. In all instances in this study, the deprivation of oxygen occurred at least four months prior to the removal of the brain. This meant that any change in the morphology of the individual cells would, in all likelihood, be corrected, or, if not, the cell would have died and complete degeneration taken place with the subsequent removal of the debris. Other investigators, notably Hurder (1952), have reported that the significant neural difference between animals exposed to anoxia and their controls was in the number of neurones per unit of cortical tissue. Although some evidence indicates the possibility of localized and diffuse lesions (Windle et al., 1944), it seemed, from what is known of the degenerative process and from the empirical data collected in an experiment similar to the present one, that the best procedure would be to prepare the tissues for cell counts. This, however, would not necessarily limit the possibilities for locating and identifying lesions and other gross anomalies.

TECHNIQUE

For analysis, four sections, free of artifacts and clearly perpendicular to the surface of the brain, were selected from each animal. These sections were as widely distributed as was possible while still complying with the above limitations.

Before any quantitative analysis was made, the sections were first observed for structural and cellular abnormalities, tumors, variations in apparent size of cells, and evidence of recent trauma.

The comparisons between the two groups regarding cell number proceeded along two lines. It was first recognized that as a consequence of oxygen deficiency, some of the cortical cells could be lost. In the adult tissue, however, the loss could be signified by a reduction in the thickness of the cerebral cortex or by a reduction in the number of neurones per unit of cortex. As the rats were in the fetal stage at the time of deprivation, both indications of reduction in the absolute number of cells were likely. Consequently, attempts were made to determine both with this material.

The thickness of the cortex was measured with an eyepiece micrometer (with a magnification of 35×) at five pairs of corresponding points on either side of the sections.

For the estimation of relative cell density, counts were made on 25 fields, randomly selected, from 4 sections. At a magnification of 450×, the counts were made as follows: using an eyepiece equipped with one short and four long pointers, the observer counted the number of pyramidal cells (as indicated by the appearance of the nucleus on the particular section at the particular level of focus) in layers III and V covered by the tips of the four long pointers. (The short pointer was used to focus the field.) It was possible, thereby, that within a single field as many as 4 or as few as 0 could be so covered by the tips. For the 25 fields, 100 cells could be so covered—or, possibly, none. The number of such "hits" was the datum for the individual *S* for intersubject and intergroup comparisons.

The reliability of this procedure was checked with material prepared in the same fashion as this but taken from animals other than those used in this experiment. These results are given in Table 27. A coefficient of concordance (Kendall's *W* [Siegel, 1956]) computed on these counts was .59 which is significant at the .05 level.

TABLE 27

RELIABILITY CHECK ON TECHNIQUE FOR CELL COUNT

Section and Slide	Count					Mean
	1	2	3	4	5	
1	18	17	14	19	16	16.8
2	15	14	13	14	16	14.4
3	18	16	19	19	17	17.8

Note.—The material for this reliability check was taken from a single animal, a normal adult female rat.

RESULTS

The qualitative examination of the tissues revealed no significant structural difference between the experimental and control animals. The size and shape of the cells were uniform in both groups of animals. The only pathology which could not be attributed to artifact was a glial module in one of the female control "dulls."

In estimating absolute cell number from the thickness of the cortex, the observers found greater within subject variability than between subject variability. The measurements differed more with the portion of the brain from which the section was taken than from one *S* to another. (In some instances, unfortunately, the sections from the control and experimental animals were not taken from corresponding portions of the brains—even though the original stipulation that the selection be through the region of the thalamus was fulfilled.) The range of thicknesses was from 50 to 110 units.

Finally, as can be noted in Table 28, no expected differences were discerned on the basis of relative cell counts. As it turned out, the counts for the experimental animals were higher than the counts for the comparable control groups. The means were: "bright" controls: 18.4; "bright" experimentals: 19.9; "dull" controls: 14.7; and "dull" experimentals: 17.6. Any statistically significant differences here would be contrary to the predicted findings.

TABLE 28

COMPARATIVE DENSITIES OF LAYERS III AND V OF CONTROL AND EXPERIMENTAL GROUPS

Group	Animal	Sex	Layer	Count	Mean III	Mean V	Mean Total
Bright Control	49	M	III	16.7	17.8	19.0	18.4
			V	14.5			
	81	M	III	19.0			
			V	23.5			
Bright Experimental	42	F	III	17.0	22.4	17.5	19.9
			V	17.0			
	36	F	III	23.0			
			V	16.2			
	48	M	III	24.2			
			V	12.5			
	85	M	III	25.2			
			V	24.2			
Dull Control	21	F	III	14.5	15.8	13.6	14.7
			V	13.7			
	39	F	III	16.1			
			V	12.6			
	43	F	III	17.0			
			V	15.5			
	20	M	III	15.5			
			V	12.8			
Dull Experimental	38	F	III	22.2	18.0	17.1	17.6
			V	21.5			
	62	F	III	19.7			
			V	14.5			
	40	M	III	19.0			
			V	17.0			
	76	M	III	17.1			
			V	17.2			
	80	M	III	17.3			
			V	17.1			
	88	M	III	12.5			
			V	15.5			

Vol. 74, No. 2 Whole No. 489, 1960

Psychological Monographs: General and Applied

SIMILARITIES IN THE STRUCTURE OF EXTREME SOCIAL ATTITUDES[1]

IRVING A. TAYLOR[2]
New York University

THE major thesis of this study, representing the outcome of several investigations of the perceptual and personal characteristics of intense "liberalism" and "conservatism," is that certain important similarities between the two have not been adequately investigated. In this respect, the present research deviates from the usual approach, which focuses upon apparent differences.

In a sense, however, this study has continuity with previous investigations. Historically, the direction of theory and research has been to demonstrate the generality and inclusiveness of social attitudes. Early studies, for example, viewed prejudice more or less as a specific and independent attitude. More recently, the tendency has been to regard it as related to the entire personality, as part of a highly generalized system, manifesting itself even in situations where social issues are not present. The evaluation of social attitudes as conceived in the present study expands the problem by viewing extreme attitudes as alternative phenotypic manifestations of the same or similar genotype.

The primary purpose of this research was to explore perceptual reactions of individuals holding opposing positions on a general social attitude continuum ranging from relatively extreme "conservatism" to extreme "liberalism," the terms referring here to the traditional "left–right" distinction. Specifically, it was our aim to compare the preceptual reactions of individuals holding opposed social attitudes with regard to closure or completion tendencies manifested in their reproductions of relatively neutral material not directly related in any apparent manner to social issues.

Perception was regarded as a fruitful approach for obtaining relevant information since neutral material could be used. This allowed for a direct comparison of underlying personality tendencies of individuals expressing opposing social attitudes without the interference of ego controls provoked by the verbal content of attitude scales.

This study of similarities revealed through perception was prompted by the consideration that the intensity and degree of a social attitude is equally important or even more important with regard to social relations than the manifest content or direction of that attitude. Moreover, perceptual behavior seems to have more to do with the generalized quality of the character structure than does the attitudinal content. It must be emphasized, however, that indications of underlying similarities do not reduce the reality and importance of overt social differences. Rather, it is the writer's intention to throw additional light on the structure of extreme social attitudes.

An operational description of the key terms used in the investigation may serve to clarify subsequent questions. The term perceptual closure is interpreted as signifying a general tendency to complete what is incompletely presented in the field of vision, presumably for purposes of effecting a more stable perception. This is measured in terms of the degree to which an individual reduces the size of openings when reproducing a figure containing several opened lines. The essential conditions involved are (*a*) a situation which is clearly incomplete, and where ambiguity and discovery are not essential factors, and (*b*) an opportunity to resolve the open situation through overt activity.

---

[1] This study received the award by the American Association for the Advancement of Science in 1957 for meritorious sociopsychological research, and is based upon a doctoral dissertation at New York University, 1954.

The author wishes to express his gratitude to Marie Jahoda for her valuable guidance and assistance, and to Richard Christie, Richard I. Evans, and G. B. Vetter for their many helpful suggestions. Appreciation is also extended to L. T. Callicutt, Chairman of the Department of Psychology of the University of Houston, and members of the Department for their cooperation and assistance with regard to subjects and the administration of tests.

[2] Now at Northeast Missouri State Teachers College, Kirksville, Mo.

To avoid confusion with other current usages, extreme "liberalism" refers here to a relatively intense ideological predisposition characterized by acceptance of minorities and by the rejection of authoritarian power groups and beliefs. The essential criteria are (*a*) a direct and positive indication of acceptance of ethnic minority groups and (*b*) a direct and positive indication of rejection of authoritarianism. The importance of obtaining direct and positive indications of both acceptance and rejection is discussed in a later portion of this monograph.

Extreme "conservatism" is conceived as a relatively intense ideological predisposition involving acceptance of authoritarian power groups and beliefs and hostile attitudes toward ethnic minorities. The essential conditions are (*a*) a direct and positive indication of acceptance of authoritarianism and (*b*) a direct and positive indication of ethnic minority rejection. The formulation, again, includes acceptance and rejection components.

Attitudes which reflect neither strong prejudices with regard to any social group nor strong beliefs with regard to any social issue are regarded as reflecting moderate or intermediate attitudes. These attitudes, therefore, lack the intensity involved in extreme ideological predispositions.

The selection of perceptual closure as a variable for investigation is prompted largely by what may be described as structural similarities suspected to underlie the behavioral pattern of extreme social groups. By a structural similarity we mean a characteristic such as the intensity of an attitude which can be expressed in apparently diverse attitudes. In many respects this formulation is similar to Klein's (1951) conception of *Anschauung* or "perceptual attitude," as indicated in the following passage:

> The term *Anschauung* is preferable to "attitude" because of the rather more narrow, well-worn connotations of the latter term in American psychologies, particularly in social psychology. Attitude usually implies a quite specific *content* and a direction toward or away from an object. In this common meaning it carries no implications of formal personality structure. But our use of it is precisely in the latter sense as a genotypic principle of control, having no ties to specific content, not necessarily related to particular conflicts or stresses, and with counterparts in all forms of cognitive behavior (p. 332).

Thus, the extreme conservative may express hostility and prejudice toward members of minority groups; the extreme liberal, similarly, may express the same degree of hostility and prejudice toward members of antidemocratic groups. Stereotypy, frequently associated with the fascist, is also apparent in the extreme "left"; "Negroes are inferior" and "Capitalists are exploiters" are both highly stereotypical. As opposed social extremists proceed farther apart ideologically, the structural similarities appear to become even more similar. Whereas both the groups and the institutions toward which the attitude is directed may differ widely, the intensity is comparable. Extreme hostility, prejudice, and stereotypy appear to be exaggerated forms of behavior, presumably having the function of reducing tensions initiated by open social situations. The tendency toward extreme social attitudes is reflected by the nature of the perceptual reproduction technique utilized in the present study.

To the extent that several studies have already supported the general hypothesis that extremely conservative or ethnocentric individuals manifest closure behavior, or closely related behavior, in a situation apparently unrelated to social issues, our contention is not new. What is novel in our research, however, is the hypothesis that this tendency is similarly discernible in the perceptual reactions of those of the "left."

The following hypotheses are specifically formulated to test the relationship between perceptual closure and extreme social attitudes.

1. There is no apparent difference in perceptual closure reaction between individuals predisposed toward extreme conservative social attitude and those predisposed toward extreme liberal social attitude.

2. Individuals predisposed toward extreme liberal social attitude manifest greater perceptual closure reaction than those exhibiting moderate social attitude.

3. Individuals predisposed toward extreme conservative social attitude manifest greater perceptual closure reaction than those exhibiting moderate social attitude.

RELEVANT LITERATURE

The Authoritarian Personality

One of the major contributions to the area of ideological predispositions has been *The Authoritarian Personality* (Adorno, Frenkel-Brunswik, Levinson, & Sanford, 1950). Early in the investigation, a scale was produced having as its purpose the measurement of ethnocentrism (E scale), designed to measure prejudice directly. In subsequent analysis, however, another scale was evolved, the California Fascism Scale (F scale), having as its purpose the measurement of implicit antidemocratic trends, or prefascistic potential, at the personality level, as distinct from manifest antidemocratic tendencies. The F scale was intended to yield an estimate of minority group prejudice as measured more directly by the E scale. The final revision of the F scale (the one used in the present study) correlated .75 with the E scale.

The major variables underlying the items of the F scale included conventionalism, authoritarian submission and aggression, anti-intraception, superstition and stereotypy, and projectivity and sex. These were regarded as central factors related to surface attitudes. These variables, according to the authors (Adorno et al., 1950) ". . . taken together, made up the basic content of the F scale. Each was regarded as a more or less central trend in the person which, in accordance with some dynamic process, expressed itself on the surface in ethnocentrism as well as in diverse psychologically related opinions and attitudes" (p. 228).

The items of the scale were constructed according to a principle of maximum indirectness with regard to overt prejudices. In the final revised scale all items were formulated in terms favorable to antidemocratic practices and beliefs. It is important to note, therefore, that a low scorer—one rejecting scale items—is "democratic" or "liberal" by implication only.

The procedures utilized by the authors of *The Authoritarian Personality* included interview, clinical, and questionnaire techniques. The primary basis for interpreting individual predispositions was derived essentially from interview protocols of extreme scorers, those tending either to agree or disagree strongly with the items of the scale. Linearity of social attitudes was, therefore, assumed, and the middle group was consequently excluded. The significance of this exclusion will be discussed shortly.

A major finding of *The Authoritarian Personality* was the identification of a well-defined syndrome characterizing high scorers. Subsequent studies have attempted to substantiate these findings as well as demonstrate further differentiations between low and high scorers. In many instances, however, the practice of excluding middles has been continued, thus perpetuating the assumption of "left–right" linearity and obscuring any curvilinear tendencies that may have been revealed in the data. In addition, most studies have either implicitly or explicitly identified low scores with "liberalism" despite the inferential basis of this interpretation and without regard to the fact that the authors of *The Authoritarian Personality* found the low scorers to be relatively more diverse as a group.

In an early study by Rokeach (1948), individuals above the median of the E scale were found to be more rigid with regard to solving *Einstellung* problems than those below. That the highly ethnocentric individual approached social problems more rigidly was apparent. The finding that this carried over to situations not directly related to social problems led Rokeach to postulate a factor of "generalized mental rigidity" in ethnocentrism. A further study by R. W. Brown (1952), in which the F scale was used in conjunction with the

Einstellung problems, revealed that differences in problem solving rigidity between low and high scorers became manifest only under conditions of ego involvement.

Rokeach has shown differences between low and high scorers in several other studies. In one study (Rokeach, 1950) the Levinson 10 item Ethnocentrism Scale was administered and subjects (*S*s) were asked to describe the way in which they thought 10 ideological terms (Buddhism, Capitalism, Fascism, etc.) were interrelated. The low quartile demonstrated the greatest amount of comprehensive organization, the middle quartiles a more isolated organization, and the high quartile a narrow one, thus indicating a linear trend. In a further study (Rokeach, 1951), *S*s were asked to define these ideological terms. Although the lowest quartile in the E scale responded more frequently with relatively abstract definitions as expected, they were also found to favor the type of "reification" usually associated with the higher scorers. Rokeach (1952) has more recently produced a scale intended to measure the "dogmatic personality" reported to be independent of the traditional "left–right" continuum.

In a study by Frenkel-Brunswik (1951), ethnic prejudice was linked with "intolerance of ambiguity." Noting an inherent ambiguity intolerance in the social responses of prejudiced children, the hypothesis was tested that this intolerance reveals itself also in neutral material. A technique was utilized in which a series of pictures were presented one at a time beginning with a clear percept of a dog which gradually transforms into a cat. The results indicated that children high in ethnic prejudice perseverated longest with their initial response as the picture series progressed. Also included in her study were cognitive materials in which similar results were obtained. She concluded, "There is some indication of a prevalence of premature reduction of ambiguous cognitive patterns to certainty in the prejudiced subjects, as revealed by a clinging to the familiar . . ." (p. 140). She further noted a close relationship between "intolerance of ambiguity" and Rokeach's "generalized mental rigidity."

Highly similar to the two studies mentioned above is the one by Kutner (1958), in which children were administered problem solving tests. Children scoring low in ethnic prejudice gave evidence of an ability for abstract conceptualization while high scoring children presented a more concrete approach to problem solving with an accompanying higher rate of failure.

The results of an experiment by Block and Block (1951) indicated that *S*s relatively high in ethnocentrism were quicker to establish autokinetic norms than relatively low scorers. This rapid establishment of frames of reference was identified with excessive ego controls by the authors although the similarity to Frenkel-Brunswik's "intolerance of ambiguity" was noted. The authors, in discussing their statistical approach, make the following somewhat dubious comment concerning the middle group:

> Although the latter analytical procedure [chi square] probably capitalizes somewhat on chance in the decision as to how many subjects are to be included among the 'Highs' and 'Lows,' it is substantially justified as a means of removing the middle range of ethnocentrism scores. It is here that discrimination is most difficult to achieve and such relationships as exist tend to be obscured (p. 309).

The middle group, however, can be considered as obscuring a relationship only if it is assumed to be linear.

In an experiment involving changes in the memory of geometric forms, Fisher (1951) has found that high scorers in ethnocentrism showed a strong tendency toward symmetry and simplification, whereas low scorers were more able to "break" levelling tendencies. Although it was hypothesized that the law of *Prägnanz* would apply to both extremes, only the lows would withstand or "break the Gestalt." Although Fisher states at the outset that ethnocentrism is considered as a continuous variable, the data only for extremes are presented with a cautious note that the results are limited to the extremes. It is not clear why the data for the middle group were not presented since this would have allowed for a direct test and demonstration of linearity, and would have eliminated the necessity for confining the results to extremes only.

In a study by Barron, reported by Frenkel-Brunswik (1954), in which an aesthetic preference test was used, it was found that ". . . subjects preferring the [simple] symmetrical figures tend to be less original, less intellectual, less sensual, less internalized, more constricted, more conservative, more ethnocentric, more conforming, and less independent" (p. 249). Many other studies using the California scales have reported differences between extreme low and high scorers. In a study by Evans (1952), prejudiced *S*s more often favored political and economic values, while unprejudiced *S*s tended more toward the aesthetic and theoretical, revealed by the Allport-Vernon Study of values. I. A. Taylor (1951) has demonstrated a slight tendency for high scorers to anticipate the correct sequence of events in a picture series test. Meer (1952) has shown that *S*s measuring high in authoritarianism report hostility toward strangers in their dreams more frequently than toward family, friends, and other ingroup members. The differences found for low authoritarian *S*s with regard to ingroups vs. outgroup hostility, however, were not significant. In a study of personality differences between volunteer and nonvolunteer *S*s, Rosen (1951) has found that the former were less subject to conventionalism, authoritarianism, and other personality characteristics presumably measured by the F scale. This was interpreted as indicating that volunteering is not entirely a function of the situation, but is related to personality characteristics.

Relying upon the evidence presented in *The Authoritarian Personality* that a fair estimate of ethnocentrism measured by the E scale could be predicted from the F scale, we have reported the results of studies using either scale. Several studies, moreover, have independently reported high correlations between the two scales. Srole (1951), however, has raised a problem concerning this relationship. In a study in which a five item F scale and an ethnic minority social distance scale were administered in conjunction with a five item Anomie scale presumably measuring psychosocial factors related to social dysfunction or disorganization, analysis of the data suggested that the F scale did not highly correlate with ethnic prejudice independently of anomic factors, and that prejudice was more closely related to anomie than to authoritarianism. Roberts, Rokeach, and McKitrick (1952), in a repeat study, however, found that prejudice and authoritarianism correlated .64 and prejudice and anomie .55, but when anomie and education were held constant, prejudice and authoritarianism correlated .48, whereas the correlation dropped to .28 between prejudice and anomie when authoritarianism and education were held constant. The writers concluded, therefore, that there is insufficient support for Srole's contention that anomie is more closely related to prejudice, and that prejudice and authoritarianism are not highly related independently, but that the presence of a low correlation between anomie and prejudice is worthy of further investigation. The interchangeability of the California scales in research, therefore, appears to be justified.

With regard to the general validity of the California scales, our first concern, Eager and Smith (1952), using a modified form of the F scale, have shown that agreement existed between summer campers' perception of the degree of authoritarian behavior shown by counselors and the scores obtained by the latter on the scale. Campbell and McCandless (1951) have reported results similar to those in *The Authoritarian Personality*. Using an independently constructed scale designed to measure racial attitudes, their findings may be interpreted as evidence for the general validity of the California results.

Christie and Garcia (1951) empirically tested the hypothetical variables underlying the items of the F scale (conventionalism, authoritarian submission, authoritarian aggression, anti-intraception, superstition and stereotypy, and projectivity and sex). The empirical clusters resulting from the analysis of interitem correlations appeared to agree fairly well with the hypothetical ones, although some items did not fall into any of the defined clusters.

Although there appears to be sufficient ground for claiming general validity for the California scales and the underlying variables, it must be emphasized that actual behavior itself is not necessarily the validating criterion of social attitudes. Thurstone (1954) has emphasized this important point:

> In studying the measurement of social attitudes, the attempt is sometimes made to validate such experiments in terms of overt behavior, but that is an error . . . . A man may be entirely consistent in what he says and in what he does about a controversial issue, and yet both of these indices may be dead wrong in reflecting his attitude . . . . Attitudes are essentially subjective experiences which may or may not conform with overt action (pp. 51-52).

The suggested criterion for validation, according to Thurstone, can be ascertained only through intimate face to face private conversation. It seems, therefore, that the above mentioned studies constitute more a measure of consistency between overt behavior and attitude scale scores than an index of the validity of the scales. The scope of the present paper does not permit a further discussion of this important point.

Our second major concern involves the meaning of low and high scores on the California scales. It was previously mentioned that the items of the scales were favorable to authoritarianism, as reflected, for example, in the F scale item, "Obedience and respect for authority are the most important virtues children should learn."

One problem presented by a scale containing items of this nature involves the general tendency for some *S*s to give positive responses independent of content. Cohn (1953), for example, reports a significant .25 correlation between the F scale and 33 items selected from the Minnesota Multiphasic Personality Inventory. Many of the items used by Cohn, however, seem to be in agreement with the underlying F scale variables. The MMPI items "At times I feel like smashing things" and "At times I have a strong urge to do something harmful or shocking" seem to be possible additional "authoritarian aggression" F scale items; and "When I take a new job, I like to be tipped off on who should be gotten next to" is suggestive of the "power and 'toughness'" variable. It is, therefore, surprising that the correlation between the F scale and the

MMPI items was not higher. The correlation may have been weakened by the inclusion of certain MMPI items which, although authoritarian in flavor, were more of the "extreme liberal" variety, as, for example, the item, "Some people are so bossy that I feel like doing the opposite of what they request, even though I know they are right." The effects of the general tendency to respond positively, therefore, is open for further investigation.

Another problem relevant to the present study regarding the construction of the California scale items involves the dubious assumption that low scorers rejecting antidemocratic beliefs are thereby necessarily "democratic" or "liberal." Whereas, a high score may reasonably reflect an inclination toward fascistic ideology (since high scores indicate agreement with the explicit content of the items), low scores merely reflect the absence of such agreement. The beliefs and attitudes imputed to the low scorers are, therefore, inferred.

It follows that when using the California scales high scorers as a group would be more homogeneous than low scorers, since the former have common grounds for agreement while the basis for low scorers' disagreement remains varied and unknown. It is not the agreement per se which produces greater homogeneity among high scorers, but rather the unmistakable conservative and authoritarian quality of the items which suggest the identity of agreers. We cannot infer, unfortunately, the reasons which prompt disagreement. There is little doubt that extreme liberality is one reason, and probably constitutes a large percentage of the low population. Reasons, however, other than extreme liberality may be influential. These may include: strong opposition to any opposing ideology while maintaining moderation in one's own beliefs; general personality traits such as the tendency to disagree; resentment against unqualified and ambiguous items; or, merely, an expression of disagreement with the test situation itself.

There are additional reasons for suspecting that a low score on the California scales does not necessarily reflect liberal predispositions. There is the finding in *The Authoritarian Personality* that low scorers as a group were more diverse than high scorers, and were categorized in several subtypes, one of which was termed the "rigid low." Other studies, moreover, have found greater similarities among high scorers than among low scorers. Rokeach (1950), for example, found more consistency among highs than among lows, and Meer (1952) did not find significant differences among the lows, although the pattern was clear for the highs.

To test the hypothesis that low scores on the F scale do not necessarily reflect extreme liberality, the investigator administered a 12 item F scale in conjunction with a modified version of the Bogardus Social Distance Scale, involving ethnic minorities, to 21 secondary school *S*s. In confirmation, it was found that whereas approximately 67% of the *S*s were in the high third of both scales, and 50% in the middle third of both scales, only 29% of the lowest scorers on the F scale were the lowest scorers on the social distance scale. Low scorers on the F scale, moreover, did not show more indication of ethnic minority acceptance than the middle F scale scorers. In fact, there was some indication of curvilinearity to the extent that the mean on the social distance scale for low F scale scorers was slightly greater than the mean on the social distance scale for middle F scale scorers. The high scorers on the F scale, on the other hand, showed more consistency in agreement with the social distance scale.

Since the above study was limited to too few *S*s, we cannot safely generalize. The implications, however, were important to the present research in which comparisons were made between groups opposed in social attitudes. It was largely for this reason that we have defined the extreme liberal, extreme conservative, and the intermediate groups on the basis of two independent criteria—those demonstrating consistency in score on both the F scale and social distance scale. This precaution increases the probability of correctly identifying the position of *S*s along a general social attitude continuum.

A third major concern, and the most important, is with regard to the assumption of linearity of social attitudes. Linearity was largely taken for granted by the investigators of *The Authoritarian Personality,* as reflected in their decision to limit interviews to extreme scorers. The exclusion of intermediate scorers was unfortunate, for it prevented a comparison between the middles and extremes. The entire research, as in the case of many subsequent investigations, was geared apparently toward differentiating lows and highs, thereby obscuring any curvilinearity that may have been present.

Hyman and Sheatsley (1954) have raised this same methodological criticism, objecting further that "The findings obtained from interviews with opposing extreme groups are generalized to the non-extreme middle groups" (p. 69), and that "Research studies have frequently revealed a curvilinear relation between social attitudes and various determinants of those attitudes, and a number of investigations have indicated that extreme individuals differ from others in the degree of organization of their sentiments and in the relation of such sentiments to personality factors" (p. 66). For somewhat different reasons, Luchins (1950) has raised a number of similar criticisms, and suggests the importance of having interviewed a sample of *S*s in the middle category as a necessary control for the linear assumption.

Actually, the important issue involved here is the implication which follows directly from the assumption of linearity. This is the contention that there are basic differences between the personalities of extreme scorers. As a consequence, the middle group becomes an unimportant link between extremes which can be justifiably eliminated. Although there are a number of alternatives to the assumption of linearity, the one that concerns us here is curvilinearity, which stresses basic similarities in the personality of extremes. The middle group then becomes the decisive factor and must be presented in relation to extremes.

Linearity and the contention of opposition of extremes is explicit in the following discussion by J. F. Brown (1936) concerning the similarities and differences between the fascist and communist dictatorships:

> The basic "purpose" of the fascist dictatorship is to assure freedom of economic locomotion to the bourgeoisie. The basic "purpose" of the communist dictatorship is to assure freedom of economic locomotion to the proletariat. . . . this variance in the class basis of the dictatorships makes them as different from each other as black from white. The similarities from the standpoint of a field-theoretical analysis are of a very superficial nature, *the differences are basic* [Italics added] (p. 410).

In an almost identical context, Shils (1954), some two decades later, discusses what appears to be a similar issue, but with the alternative implication of basic similarities between extremes and the superficiality of differences:

> Obviously a Fascist who says that the Jews have monopolized almost all the important posts in the Government is concretely different from the Bolshevik who asserts that the small circle of big business men control not only the economic life but the intellectual, political, and religious life of the country. Concretely these two views are very different—one we know to be the usual paranoid anti-Semitism—the other sounds like a somewhat crude social science proposition in which many intellectuals in the West believe. Yet looked at from another point of view, *they are strikingly similar* [Italics added]. Both aver that a small group has with doubtful legitimacy concentrated the power of the country in their hands (p. 32).

Thus, with striking contrast, the same phenotypic situation is considered genotypically different by Brown, but similar by Shils.

In another place, Brown asserts that there are ". . . basic differences [between the fascist and communist state] in economic freedom for the different classes and the different positions taken with regard to nationalism, racial minorities, education, the arts, position of women, internationalism, etc." (p. 411). Since the publication of Brown's book, however, many of these differences are no longer apparent. The development of communistic ideology has increasingly come to resemble fascism in recent years, revealing similarities previously obscured. Hoffer (1951), thus, points out an important similarity in an area believed by Brown to involve basic differences: "The Bolshevik and Nazi revolutions are also full-blown nationalist movements. The Nazi revolution had been so from the beginning, while the nationalism of the Bolsheviks was a late development" (p. 18).

When comparisons are made between extremes in which the California scales are *not* employed, and when an examination is also made of the middle group, the results frequently indicate similarities rather than differences. Vetter (1930), in studying groups along a reactionary–radical continuum, has reported a curvilinear relationship between extremes with regard to introversion. Reactionary men scored 16.1 and radicals 18.5, but intermediate conservatives scored only 14.8 on the Laird Personal Inventory. He also found that while radicals had an average income of $7,100 and reactionaries of $7,700, the intermediate group had an income of $10,500. One of the results of the present research clearly supports Vetter's findings. Allport (1937) has suggested that "a liberal is not merely a mild radical; in many respects he has a wholly different outlook. Nor is a reactionary merely an ultra-conservative, for he resembles the radical in many respects: he too favors change (albeit towards the old rather than the new); he too is discontented, and . . . seems to be as excitable and maladjusted as the chronic radical" (p. 431). Dombrose and Levinson (1950), indeed, have recently confirmed Allport's hypothesis that "a liberal is not merely a mild radical," by showing that the low quartile scorers on the E scale prefer militant measures in distinction to the middle low quartile scorers' preference for pacifist programs. The writers conclude that ". . . the greater the intensity of support for democratic ideology, the greater the tendency to prefer militant forms of social action" (p. 107), and that ". . . the militant low would attempt to change whatever social patterns . . . he regards as responsible for the growth . . . of antidemocratic groups and ideologies" (p. 108). The findings of Dombrose and Levinson support the contention that the extreme low scorer on the California scales favors a form of behavior frequently associated with the high scorers.

A study by Haimowitz and Haimowitz (1950) supports the above contention. Twenty-four *S*s invited to engage in a therapeutic training session were administered the Bogardus Social Distance Scale involving ethnic minority groups. Several weeks later, the scale was readministered. Also included in the second social distance scale was a list of groups suspected to be at the focus of the extreme liberal's hostility. The results indicated that *S*s most friendly to ethnic minorities were most hostile to antiminority groups, a finding which seems to be in opposition to one of the results of the present study. Of the 24 *S*s used, only the middle group showed any change during the training period, and, with regard to extreme individuals, the authors state that modification of ". . . such [social] attitudes is more difficult in these people than in others who hold to attitudes of either friendliness or hostility with less intensity. Thus, the variable which makes for modifiability of attitudes *is not so much the factor of directionality of attitudes, as it is of intensity*" [italics added] (p. 236). The writers continue that ". . . persons who do not exclude ethnic groups may be 'anti-democratic' in their generalized response to the world" (p. 237). On the basis of several other pertinent findings, the authors conclude, ". . . the two extreme groups were more similar to each other than they were to the intermediate group" (p. 240). This conclusion is clearly in agreement with the one based on our results.

An affinity between the extreme groups has been strongly asserted by the historian Schlesinger (1949), who states, ". . . from one viewpoint the similarities [between the communists and fascists] are vastly more overpowering and significant than the differences . . . the passage from the extreme left to the extreme right and back has been fast and easy" (pp. 59–60). Hoffer (1951), in a discussion of the interchangeability of mass movements, similarly states that "When people are ripe for a mass movement, they are usually ripe for any effective movement, and not solely for one with a particular doctrine or program. In pre-Hitlerian Germany it was often a tossup whether a restless youth would join the Communists or the Nazis" (p. 16).

Shils (1954), in close agreement with the above, states that ". . . there is almost always a significant number of low scorers whose deeper dispositions are closer to the high scorer than to the low scorer. In conformity with the preconceived idea that authoritarianism is a characteristic of the Right and the corresponding notion that there is no authoritarianism on the 'left,' there is no analysis of these deviant low scorers" (p. 32).

Krugman (1952) has indeed investigated a group that may be considered as "deviant low scorers." Ex-members of the Communist Party were interviewed, and the author reports many instances of behavior bearing a striking similarity to authoritarian practices associated with the "right."

There appears, however, to be an inconsistency inherent in the notion of "deviant low scorers," as exemplified in the above passage by Shils in which he states that some low scorers are "closer to the High Scorer than to the Low Scorer." There seems to be the implication that there is a subvariety of renegade low scorer similar to the high scorer, and that the other low scorers are the genuine liberals having personality characteristics basically different from the high scorer. This implication is also apparent in the following statement by Frenkel-Brunswik (1954):

> Such concepts as that of the 'rigid low,' introduced in our book . . . amply testify to the fact that we have not lost sight of the possibility that there may be curvilinear relationships among the various attitudes or traits of personality and that certain sub-varieties of the non-ethnocentric personality likewise exhibit the closeness of opposites that is bound to accompany intolerance or cognitive ambiguity or emotional ambivalence (p. 258).

This inconsistency can perhaps be resolved by considering only the "rigid lows" as the extreme liberals, and the remaining low scorers as individuals strongly disagreeing with the items of the California scales for reasons other than extreme liberality. This interpretation is congruent with our previous discussion of the meaning of a low score.

Frenkel-Brunswik (1954) has also stated that "In our view, the susceptibility to ethnic prejudice itself is but another case of 'alternative' behavior" (p. 261), and, at another time, that "The relatively incidental 'choice' between different alternative manifestations of the *same underlying tendency* [italics added] is most crucial for the development of personality. Thus, whether a person is ethnocentric or liberal is closely related to whether his underlying aggression is accepted or not . . ." (pp. 265-266). Allport and Kramer (1946) imply a similar thought in suggesting that " . . . the feeling of being a victim of prejudice on the part of the [Jewish] subjects may lead to sympathy or antipathy toward the Negro" (p. 29).

There appears to be a conflicting tendency discernible in the literature to consider, on the one hand, only the "rigid low" as basically similar to the extreme conservative, and, on the other hand, to regard all extreme liberals as essentially similar. The proposed suggestion to limit the category "extreme liberal" to the "rigid low" scorer only has the benefit of resolving this discrepancy, and also has the advantage of allowing for an understanding of the manifest differences as instances of one genotype.

In an exploratory study by B. Taylor (1952), which used an instrument for measuring closure similar to the one used in the present research, a significant positive relationship between closure and the F scale was found. The median division was used to differentiate low and high scorers. When *S*s were divided into quarters, however, and a mean perceptual closure score was computed for each group, a tendency toward curvilinearity became apparent.

In a second exploratory study in which a social distance scale involving ethnic minority groups was administered to secondary school children in conjunction with the F scale and a perceptual closure scale, *S*s were selected scoring in the low, middle, and high third on both social attitude scales. Those falling in the extreme categories exhibited the greatest closure tendencies, as predicted. It appeared, therefore, that perceptual closure tendencies occur for both extremes when independent measures are used conjointly to estimate the intensity of social attitudes. The sample, however, was too small for confident generalization.

Perceptual Closure

Reviewing the relevant literature on perception, particularly on perceptual closure, we find that the term has been used in widely different contexts. Ever since its introduction by the Gestalt psychologists its usage has been extended beyond the original limitation to perception. The following statement by Hochberg and Gleitman (1950), in discussing a nonseparatist theory of motivation, exemplifies this pervasive extension of the term:

> Such a theory of motivation would consider all motivational phenomena as manifestations of equilibratory processes occurring in the psychophysical field. From this point of view, the formally simplest motivational phenomenon might be—not the rat running towards the food, not the infant establishing object-relations with its mother—but the much more prosaic case of an incomplete circle tending to close (p. 188).

As a basic concept of Gestalt theory, the term generally refers to an organizational process in which objects are perceived as complete or as "good" a figure as possible under the conditions obtaining. The term, however, as pointed out by Bobbitt (1942), has been expanded to include " . . . any psychological phenomena in which a condition of incompleteness, either in the stimulus field or in some phase of the organism's activities, creates a tendency to overcome this incompleteness by perceptual reorganization, by a combination of two or more different experiences, or by some overt activity of the organism" (p. 274). The generality of the term is similarly indicated by Drever's (1953) definition of closure as a principle ". . . describing the process by which percepts, memories, actions, etc., attain stability, viz. the subjective closing of gaps, or completion of incomplete forms, so as to constitute wholes" (p. 40).

Although the early Gestalt formulation of perceptual closure regards the tendency to close gaps as revealing a fundamental structural principle of brain dynamics, other investigators have tended to emphasize the role of past experience in determining the perception of a figure. Bruner (1951), for example, states that ". . . the process of closure . . . is an abstraction based upon a sampling of various situations in which different expectancies have made different trace systems available for communication with input stimulus information" (p. 310).

The general characteristics of completion tendencies have been carefully investigated and described in an early study by Gibson (1929). Using relatively meaningless geometric forms containing gaps or breaks to be reproduced by *S*s Gibson showed that completing tendencies occurred for the group as a whole, and he concluded that "Completion occurs when the stimulus figure is apprehended as a single form, despite the gaps, and when the reproduction is drawn with a continuous contour" (p. 27). Gibson is also inclined to relate his results to previous experience rather than to laws of configurations. Gibson also pointed out the important feature of differences between *S*s with regard to degree of closure. Only 15% of the group closed gaps completely. Similar results were obtained in earlier studies by Hemstead (1901) and Meakin (1903).

Thurstone (1944), in an extensive analysis of the factors involved in perception, has isolated an important one which he has called "closure," described as being concerned with the facility, rather than acuity, in perceptual closure. An examination of the tests found to be loaded with this factor " . . . seem to represent the ability to form a perceptual closure against some distraction. They also represent the ability to hold a closure against distraction" (p. 101). Thurstone further points out three specific usages of the term: (*a*) when what appears to be unrelated and unorganized suddenly appears as a meaningfully related whole; (*b*) when a part of the presented items are suddenly perceived as a unity and the remaining parts are perceptually discarded; and (*c*) when a good configuration has to be destroyed in order to discover a different configuration (p. 20). It would seem that these usages are more closely related to other concepts than to closure: the first, to "insight"; the second, to "field and ground"; and the third, to "flexibility." Moreover, these usages deviate far from the original concept of perceptual closure. Further, all three usages involve a principle of discovery and ambiguity which is not an essential component of the process.

Although the above usages of the term differ largely from the one used here, the following statement by Thurstone (1944) clearly expresses an important proposition underlying this study:

> The fundamental hypothesis involved here is that the dynamics of perception, and of other restricted functions, are not isolated and that these several functions are so related that some characteristics of the person as a whole might be inferred from the dynamics of one of these functions. In these days when we insist so frequently on the interdependence of all aspects of personality, it would be difficult to maintain that any of these functions, such as perception, is isolated from the rest of the dynamical system that constitutes the person (p. 3).

Roseman (1952), in fact, has related personality organization to perceptual closure measured by a technique following Thurstone's first usage of the term (when what appear to be unrelated items suddenly appear as a meaningfully related whole). Psychiatric patients and normal *S*s were given groups of parts that could be formed into single geometric figures. Of the 29 closure items the normal *S*s succeeded in completing slightly fewer than 22, while the psychiatric patients completed slightly fewer than 18. The difference between the two means was significant.

Relevant also is the study by Frenkel-Brunswik (1951), where children were presented with a series of cards showing a gradual transformation from a dog to a cat, a technique closely exemplifying the third usage of closure suggested by Thurstone (where a good configuration has to be destroyed in order to discover a different one). The present study shares in common with the above study the belief that there is a relationship between the closure of open figures and the "closure" of open social issues involved in extreme social attitudes. The hypothesis here, however, is that the extremes—both the "left" and the "right"—would demonstrate closure tendencies.

DEVELOPMENT OF THE EXPERIMENTAL MATERIAL, DESIGN, AND THE SELECTION OF SUBJECTS

An estimate of antidemocratic or authoritarian tendency was obtained with the use of the 30 item revised California Fascism Scale (F scale). The scale has a reported reliability coefficient of .90 and correlates approximately .75 with the California Ethnocentrism Scale (E scale). Likert-type scaling is employed, and a score reflects the degree to which an individual accepts or rejects implicit antidemocratic beliefs. The F scale is asymmetrical; all of the items are stated in antidemocratic form (Adorno et al., 1950). A copy of the F scale is included in Appendix A.

Additional measures of social attitudes were obtained through the use of a social distance scale containing two subscales, a minority subscale with 10 ethnic minority groups and an authority subscale with 10 authority groups (see Appendix A). Measures of overt hostility or friendliness toward the two groups were obtained simultaneously.

The social distance scale is a questionnaire in which the individual indicates the degree to which he will accept members of various groups. "Social distance" has been defined by Bogardus (1933) as "the degree of sympathetic understanding that exists between two persons or between a person and a group" (p. 265). The score is the number of checked inclusions of groups for the following situations:

1. To close kinship by marriage
2. To my club as personal chums
3. To my street as neighbors
4. To employment in my occupation in my country
5. To citizenship in my country
6. As visitors to my country
7. Would exclude from my country

The score reflects policy orientations rather than feelings towards groups. The final category, "Would exclude from my country," was used as a check against the preceding acceptance categories, i.e., a *S* checked either the categories 1 to 6, or only category 7.

The ethnic minority groups included Negroes, Turks, Mexicans, Puerto Ricans, and Hindus. Representative authoritarian or antiminority groups suspected to be at the focus of the extreme liberals' hostility were Fascists, Nazis, Reactionaries, Ku Klux Klan, and Totalitarians. The minority and authority groups were presented in random order. Although the social distance scale has several inherent weaknesses, it was considered feasible as a supplementary indicator of social attitudes.

Two scales were developed for measuring perceptual closure (see Appendix B). Previous results with the use of several similar instruments have indicated that meaningful figures afford more reliable scores than relatively less meaningful geometric forms. After preliminary pilot studies, the outlines of a house and a boy were selected. Each figure was constructed of several lines containing small gaps or breaks. The figures were presented to the *S*s with instructions to reproduce them "just as they are seen."

Below the figures, all terminal points were provided for establishing the general outline upon which the figure was to be reproduced. Holding the size of the lines of the reproduced figure equal to those of the figure provided a standard basis for measuring the size of the reproduced gaps, thus avoiding the necessity for using ratios. Perceptual closure tendency is reflected by the mean size of the gaps in the reproductions. The size of all the gaps in the standard was 26 mm. The lower the score, therefore, the greater the closure tendency.

In order to test the hypothesis that the scores obtained on the "house" and "boy" scales reflected perceptual closure, or the tendency to complete openings in a *single* configuration, rather than mere completion tendencies, chance habits of writing, etc., the design scale was constructed as a control, in which completion tendencies between several figures, rather than one, could be measured (see Appendix B). The size of the gaps in the design scale was also 26 mm.

On the second sheet of the F scale, information was requested about the age, sex, religion, major, education, and estimated yearly income of each *S*. With the exception of age and academic major, the information was obtained in the form of check lists. *S*s were allowed to remain anonymous.

The entire population studied was taken from classes at the University of Houston during the summer of 1953, and numbered 253. The *S*s were selected from classes in psychology, education, and English. Table 1 summarizes the population characteristics of the group. Age ranged from 17 to 61, with a mean of 27.46 and a standard deviation of 8.81. The median age of 24.43 is a more appropriate estimate since age was not normally distributed. *S*s were heavily concentrated at the lower age intervals. Since the group was comprised of many teachers and other professional people attending summer sessions, the average age was somewhat larger than usual. The ratio of men to women was approximately 3:2. Nearly 80% of the group were Protestant, 17% were Catholic, and 3% were Jewish.

Since not all students were expected to be gainfully employed, an estimate of the parents' yearly income was requested for determining the *S*s' approximate economic level. The mean yearly income reported was $5,600, with a standard deviation of $1,920. Approximately 7% did not provide this information.

Table 1 also indicates that the number of students increases with the level of education, with graduate students comprising more than one quarter of the population. All levels of education, however, are well represented. The average level is somewhat above Junior.

The population was heterogeneous with regard to academic major. Practically all areas of study were well represented with business majors comprising the largest group (21%) and natural science majors, the smallest (6%).

Twelve groups were tested separately. An attempt was made to duplicate the procedures as fully as possible in each instance.

TABLE 1

Population Characteristics of the Groups Studied[a]
(253 *S*s)

Population Characteristic	*N*	%
Age		
55–64	2	.8
45–54	14	5.5
35–44	31	12.3
25–34	77	30.4
15–24	127	50.2
Unknown	2	.8
Parental income		
$9000–10000	44	17.4
7000–8000	30	11.9
5000–6000	72	28.5
3000–4000	74	29.2
1000–2000	16	6.3
Unknown	17	6.7
Religion		
Protestant	195	77.1
Catholic	43	17.0
Jewish	7	2.8
Unknown	8	3.1
Education		
Graduate	67	26.5
Senior	58	22.9
Junior	46	18.2
Sophomore	44	17.4
Freshman	31	12.2
Unknown	7	2.8
Major		
Business	53	21.0
Education	52	20.6
Liberal Arts	40	15.8
Social Arts	34	13.4
Biological Science	20	7.9
Technology	19	7.5
Natural Science	15	5.9
Unknown	20	7.9
Sex		
Men	146	57.7
Women	103	40.7
Unknown	4	1.6

[a] Combined from 12 psychology, education, and English courses at the University of Houston.

The initial instructions were as follows:

> This morning (afternoon, evening) I would like you to take a few public opinion questionnaires. In answering the questions it is important that they reflect your own feelings. If you are not certain about some of these issues, your first impression is probably the best. There are, of course, no right or wrong answers. Please do not put your name on any of the papers, since we are only interested in the opinions of the group as a whole. Use only the mechanical pencils that will be distributed.

It is not known whether this anonymity helps increase validity, but it does aid in dispelling tension and suspicion. Standard mechanical pencils were used to increase the comparability of reproductions on the perceptual closure scales by eliminating certain unnecessary differences due to the writing implement.

The F scale was distributed with an indication that no one was to begin until instructions were read aloud. During the administration of the questionnaire, answers to most queries concerning definitions and meanings were to the effect that responses should be in accordance with one's own interpretations.

Essentially, the same procedure was used in the administration of the social distance scale. The *S*s were requested to hold all papers until the completion of the entire battery.

The perceptual closure scales and the design scale were distributed together with the following instructions:

> This part of the study is not a public opinion questionnaire; it is another kind of test. In each of the three sheets, you are to copy the rest of the figure in the space below using the dots as guides. Notice that parts of the figure are already copied below. As soon as you complete the first, go on to the next. Do them in the order given ("House," "Boy," "Design").
>
> Work accurately and quickly. Draw free hand lines as straight as you can with the mechanical pencil. It is very important that you copy the rest of the figures by eye only.

The instructions were intended to produce a subjective, rather than an objective, set. Fortunately, there were few questions concerning the purpose of these scales, which was probably due to the interest aroused by the novelty. The scales were completed within a single one-hour class session.

The item mean score for the 253 *S*s on the F scale was 3.63, with a standard deviation of .82, which is below the F scale's average of 4.00.

One important requirement of a social distance scale is that acceptance increase with progressive levels of the scale. If, for example, a *S* indicated that he would accept Puerto Ricans "to close kinship by marriage," then Puerto Ricans should be accepted on all preceding levels (except for the last level, "Would exclude from my country"). The degree to which progressive increase in acceptance is actually obtained is indicated by the frequency of pattern reversals, e.g., accepting members of an ethnic group to one's club as personal chums, but not to one's street as neighbors. There were 277 item reversals out of a total of 2,450 possible responses (since 245 *S*s responded to 10 items) or 11.3% pattern reversals. This percentage of pattern reversal was somewhat higher than expected and probably decreased the reliability of social distance scores to some extent.

Table 2 shows the median and rank position for each of the groups contained in the social distance subscales. The five most accepted were the minority groups, and the five least accepted were the authority groups. The division of groups into minorities and authorities, therefore, was psychologically meaningful. The median acceptance for minority groups as a whole was somewhat beyond level 3 ("To my street as neighbors") while the median for the authority groups as a whole was at level 6 ("As visitors to my country"). Puerto Ricans were accepted "To my club as personal chums," while the Nazis and Ku Klux Klan would be excluded from the country. Negroes were the least accepted of the minority groups and Reactionaries were the most accepted of the authority groups. There were slightly more responses made to the minority than the authority items. The greatest amount of variability occurred for Reactionaries, which may have been due to the fact that many *S*s did not know the meaning of this term. The least amount of variability occurred for the Negroes. Most of the *S*s would neither accept nor reject

TABLE 2

Median Preference and Rank Order for Minority and Authority Items On the Social Distance Scale[a]

Social Distance Scale									
Minority Subscale					Authority Subscale				
Item	No. of Responses	*Md.*	Q	Rank	Item	No. of Responses	*Md.*	Q	Rank
Puerto Ricans	242	2.27	1.24	1	Reactionaries	231	5.11	2.00	6
Turks	237	2.74	1.47	2	Totalitarians	232	6.12	1.08	7
Mexicans	243	2.83	1.35	3	Fascists	241	6.46	.63	8
Hindus	241	3.44	1.80	4	Nazis	240	6.52	.69	9.5
Negroes	240	4.06	.57	5	Ku Klux Klan	238	6.52	1.58	9.5
All items	1203	3.30	1.31		All items	1182	6.21	1.20	

[a] Medians interpreted as follows:
1. To close kinship by marriage
2. To my club as personal chums
3. To my street as neighbors
4. To employment in my occupation in my country
5. To citizenship in my country
6. As visitors in my country
7. Would exclude from my country

Negroes beyond or below "To employment in my occupation in my country." The distribution of scores obtained on the authority subscale was extremely skewed, corresponding to a "J" curve; scores for the minority subscale, however, were normally distributed. The scale, although suitable for the minority items, apparently did not contain enough levels for the authority items at the rejection end.

The closure scales were designed to measure completion tendencies under closure conditions, i.e., when an interrupted line is perceived as forming the outline of a single configuration. The combined systematic error for the group, shown in Table 3, was −4.43 mm., or a 17.0% underestimation of the size of the gaps. Completion tendencies on the Boy scale was 1.12 mm. greater than for the House scale. The difference is highly significant ($t = 4.549$). Ten *S*s in the case of the house scale, and six *S*s with regard to the boy scale did not reproduce the configurations. The reliability coefficient for the house scale was .85; for the boy scale it was .90.

One of the major purposes of the design scale was to test the assumption that closure tendencies occur only when the interrupted outline is perceived as a single configuration. Table 3 also indicates the results for the design scale. The systematic error for the group as a whole was only .28 mm., or 1.1% overestimation. There is a 4.71 mm. difference between the combined mean of the closure scales and the design scale mean which is statistically significant beyond the 1% level. This result is interpreted as supporting the above assumption.

TABLE 3

Means and Systematic Errors in the Reproductions of Standard Openings in the Perception Scales[a]
(Standard opening: 26 mm.)

	Closure Scales			Design Scale
Item	House	Boy	Both House and Boy	
N	243	247	242	251
M	22.13	21.01	21.57	26.28
SD	4.48	4.36	4.04	4.90
Error	−3.87	−4.99	−4.43	.28
% error	14.9	19.2	17.0	1.1

[a] Errors are differences in mm. between means and standard.

It was expected that the F scale would be consistently related to the minority and authority social distance subscales. A Pearson correlation coefficient of .32 (standard error .06) was obtained between the F scale and the minority subscale, significant at the 1% level of confidence. The correlation at first appears to be smaller than might be expected since social attitudes are known to be highly generalized. The size of the correlation, however, becomes understandable if it is noted that the two scales measure social attitudes at two distinct levels (authoritarianism at an implicit, central level and social distance at an explicit, peripheral level). Even more important, both measure opposite aspects of social attitudes; extreme "conservatism" is measured directly by the F scale, and extreme "liberalism" directly by the minority subscale.

An unexpected negative correlation was obtained between the F scale and the authority social distance subscale. Examination indicated a highly skewed distribution of scores for the authority subscale which prevented the direct application of a Pearson coefficient of correlation. A contingency coefficient of −.23 was obtained. The corrected coefficient, comparable to a Pearson correlation, was −.33.

Table 4 indicates this negative relationship between the F scale and the authority subscale. The division of scores into low and high categories was made at the medians of the respective scales. A chi square of 12.955, significant at better than the 1% level of confidence, indicated that there was a tendency for *S*s scoring high on the F scale to reject authority groups to a greater extent than low scorers.

TABLE 4

THE RELATIONSHIP BETWEEN THE CALIFORNIA FASCISM SCALE AND THE AUTHORITY SUBSCALE OF THE SOCIAL DISTANCE SCALE
(Median division of *S*s)

Authority Subscale	F Scale		
	Low	High	Both Groups
Acceptance	78	49	127
Rejection	44	71	115
Both Groups	122	120	242

The correlation between the two subscales of the social distance scales was found to be positive. It was implicitly assumed in this study that extreme liberals would accept ethnic minority groups and reject authority groups, and that conservatives would reverse this order of preference. A contingency coefficient of .25 or a corrected coefficient of .37 was obtained between the subscales. The relationship is shown in Table 5 where the median point is again used to differentiate low and high scores. A chi square of 15.465, significant at better than the 1% level of confidence, indicates that *S*s tend either to accept both minority and authority groups, or reject both. Social distance, therefore, appears to be generalized, i.e., independent of specific groups.

TABLE 5

THE RELATIONSHIP BETWEEN THE SUBSCALES OF THE SOCIAL DISTANCE SCALE
(Median division of *S*s)

Authority Subscale	Minority Subscale		
	Rejection	Acceptance	Both Groups
Acceptance	42	85	127
Rejection	67	48	115
Both Groups	109	133	242

The Pearson correlation coefficient between the two closure scales was .62 (standard error, .04), which when corrected for attenuation becomes .71. The scores on both scales were combined into a single measure of closure. The combined closure scores correlated —.01 with the scores on the design scale (standard error of the correlation .07), which was not significantly different from a zero correlation. The zero correlation between the closure scales and the design scale indicates that the processes involved in reproducing the size of gaps within a single configuration is different from those processes involved in reproducing the size of gaps between nonunitary figures.

It was initially intended that the selection of samples corresponding to extreme and intermediate social attitude categories would be determined on the basis of the F scale and both subscales of the social distance scale. Extreme liberals were originally defined as rejecting authoritarian beliefs, accepting ethnic minorities, and rejecting authoritarian groups; extreme conservatives, as accepting authoritarian beliefs, rejecting ethnic minorities, and accepting authoritarian groups; and the intermediate groups, as moderate with regard to authoritarian beliefs and social distance. The social attitude categories of "low," "middle," and "high" were determined on the basis of quartile divisions, although the "middle" group was the middle 25% on all three scales. The middle quarter was defined as the combined upper half of the middle-low quarter and the lower half of the middle-high quarter. Eliminating the outer half of the two middle quarters served the purpose of producing a distinct gap between the low, middle, and high categories. Using this sampling procedure, it turned out that of the 253 *S*s only two *S*s could be described as "low," six as "middle," and one as "high." This result was due to the unanticipated negative correlation between F scale scores and authority subscale scores, and also to the positive correlation between the authority and minority subscales. It became necessary, therefore, to modify the original sampling procedure.

Since the minority subscale correlated consistently with the F scale, the best alternative was to determine social attitude categories from these two, and to eliminate the authority subscale. This modification was not serious since it was the original purpose to obtain additional direct measures of liberality provided by the minority, rather than the authority, subscale. Utilizing this modified sampling procedure, a sufficient number of *S*s scored in the consistent social attitude categories for statistical purposes. This sample, derived from the consistent quarters of the F scale and minority subscale, will be discussed subsequently.

The Authoritarian Sample

The numbers and percentages of *S*s scoring in each quarter of the F scale is shown in Table 6. Both the low and high categories each comprise approximately 24% of the population, and the middle quarter, 26.5% (not shown in Table 6). The authoritarian sample, therefore, numbered 183, or three-quarters of the entire population.

TABLE 6

NUMBERS AND PERCENTAGES OF *S*s IN THE F SCALE AND MINORITY SUBSCALES
(Division of *S*s by quartile)

Minority Subscale	F Scale									
	Low		Low-middle		High-middle		High		All Groups	
	N	%	*N*	%	*N*	%	*N*	%	*N*	%
High (rejection)	8	3.3	12	4.9	17	6.9	22	9.0	59	24.1
High-middle	10	4.1	13	5.3	16	6.5	14	5.7	53	21.6
Low-middle	12	4.9	27	11.0	17	6.9	11	4.5	67	27.4
Low (acceptance)	28	11.4	13	5.3	12	4.9	13	5.3	66	26.9
All Groups	58	23.7	65	26.5	62	25.3	60	24.5	245	100.0

The Social Distance Sample

The numbers and percentages of *S*s scoring in each quarter of the minority subscale is also shown in Table 6. There were approximately 27% in the low category, 24% in the high category, and 26.1% in the middle quarter (not shown in Table 6). The social distance sample numbered 189, or slightly more than three-quarters of the entire population.

The Consistent Sample

The consistent sample included *S*s scoring in the same quarters of the F scale and the minority subscale, shown in Table 6. There were 28 consistent low scorers (those scoring in the low quarters on the two scales), or approximately 11% of the population. The 22 *S*s in the high quarters, the consistent highs, comprise 9% of the population. The consistent middles (not shown in Table 6) numbered 19, or 7.8% of the population. There was a total of 69, or 27.3% of the entire population, in the consistent sample.

It was expected that there would be a larger percentage of *S*s in the consistent high category than in the consistent low. This was not borne out, as Table 6 indicates. Actually, the reverse occurred although the difference between the percentages is not significant.

TABLE 7

Score Limits and Numbers of Lows, Middles, and Highs in the Social Attitude Samples[a]

Social Attitude Category	Social Attitude Sample				
	Authoritarian Sample	Social Distance Sample	Consistent Sample		
			F Scale		Minority Subscale
Low					
Upper limit	3.00	2.2	3.00		2.2
Lower limit	1.70	1.0	1.70		1.0
N	58	66		22	
Middle					
Upper limit	3.93	3.8	3.93		3.8
Lower limit	3.40	2.8	3.40		2.8
N	65	64		19	
High					
Upper limit	5.43	7.0	5.43		7.0
Lower limit	4.23	4.4	4.23		4.4
N	60	59		28	
All groups					
N	183	189		69	
%[b]	74.7	77.1		27.3	

[a] Scores were interpreted as follows:

Authoritarian Sample (F scale)

1. Strong opposition, disagreement
2. Moderate opposition, disagreement
3. Slight opposition, disagreement
4. (Neither opposition nor support)
5. Slight support, agreement
6. Moderate support, agreement
7. Strong support, agreement

Social Distance Sample (Minority subscale)

1. To close kinship by marriage
2. To my club as personal chums
3. To my street as neighbors
4. To employment in my occupation in my country
5. To citizenship in my country
6. As visitors to my country
7. Would exclude from my country

Consistent Sample

(Corresponding quarters of the F scale and minority subscale).

[b] Computed on the basis of the entire population of 253.

Although the low, middle, and high social attitude categories in the authoritarian, social distance, and consistent samples are relative to the actual distribution of scores obtained, Table 7 shows a favorable correspondence between the ranges of scores for each category with the actual meanings of these scores. The categories of the consistent sample, derived from comparable quarters on the F scale and social distance scale, can be interpreted from Table 7 as follows:

The consistent low. Ss in this category exhibited moderate or slight opposition to authorization beliefs on the F scale, and accepted ethnic minorities, on the minority subscale, to kinship by marriage and to club as personal chums.

The consistent middle. Ss in this category exhibited slight opposition, or neither opposition nor support of authoritarian beliefs on the F scale, and accepted ethnic minorities, on the minority subscale, as neighbors and to employment in one's occupation.

The consistent high. Ss in the consistent high category exhibited neither opposition nor support of authoritarian beliefs on the F scale, and their response to ethnic groups, on the minority subscale, ranged from acceptance in one's occupation to exclusion from one's country.

PERCEPTION AND POPULATION SIMILARITIES IN EXTREME SOCIAL ATTITUDES

Having thus established categories of social attitudes and a single perceptual closure dimension, it was possible to test directly the hypotheses of this study. Table 8 summarizes the results for all three samples.

TABLE 8

DIFFERENCES IN MEAN PERCEPTUAL CLOSURE BETWEEN LOWS, MIDDLES, AND HIGHS OF THE SOCIAL ATTITUDE SAMPLES
(Standard opening: 26 mm.)

Sample	Social Attitude Category			Difference		
	Low	Middle	High	Low-High	Low-Middle	High-Middle
Authoritarian						
N	60	65	59			
M	21.77	22.36	20.62			
SD	3.20	3.35	5.40			
Error[a]	−4.23	−3.64	−5.38			
% error	16.3	14.0	20.7			
Difference				1.15	.59	1.74
t				1.400	.995	2.155*
Social Distance						
N	64	59	56			
M	22.10	21.15	20.63			
SD	3.26	5.32	3.62			
Error[a]	−3.90	−4.85	−5.37			
% error	15.0	18.7	20.7			
Difference				1.47	.95	.52
t				2.333*	1.196	.610
Consistent						
N	26	18	20			
M	20.80	22.86	19.95			
SD	2.28	3.67	3.36			
Error[a]	−5.20	−3.14	−6.05			
% error	20.0	12.1	23.3			
Difference				.85	2.05	2.91
t				1.000	2.232*	2.476*

* Significant at .05 level.

[a] Differences between three place means and standard in all tables.

In the authoritarian sample, the perceptual closure means for lows, middles, and highs indicated a general tendency toward curvilinearity. The highs exhibited 20.7% error in underestimation, the lows 16.3%, while the middles reduced the size of the openings by 14.0%. The highs, therefore, showed a greater closure trend than the lows, but the middles showed the least closure. It is important to note that had the middles been excluded, the general appearance of the tendency would have fostered the more familiar interpretation of linearity. Only the difference between the means of the authoritarian middles and highs is significant. The obtained *t* of 2.155 is significant at the 5% level of confidence.

The social distance sample revealed a linear relationship between the minority subscale and the closure scale as indicated in Table 8. The percentages of error increase with hostility toward the ethnic minorities. Only the difference in the perceptual closure means of the social distance lows and highs is significant (at the 5% level of confidence). If the study were limited to the minority subscale, it would appear necessary again to interpret these results in accordance with the general contention that lows are different from highs in perceptual closure reactions.

The consistent sample, however, as indicated in Table 8 showed an unmistakable curvilinear tendency. The percentages of error in underestimation for the consistent lows and highs were 20.0 and 23.3%, respectively, while for the consistent middles the reduction was only 12.1%. The extreme categories, therefore, exhibited a greater closure tendency than the middle.

Testing the difference between the means of the consistent lows and highs, a *t* of 1.000 was obtained. Since the difference is not statistically significant, this confirms our first hypothesis which stated:

> There will be no apparent difference in perceptual closure reaction between individuals predisposed toward extreme conservative social attitude and those predisposed toward extreme liberal social attitude.

Testing the difference between the means of the consistent lows and middles, a *t* of 2.232 was obtained. The difference is statistically significant at the 5% level of confidence, thus confirming the second hypothesis which stated:

> Individuals predisposed toward extreme liberal social attitude will manifest greater perceptual closure reaction than those exhibiting moderate social attitude.

In a similar manner, the difference between the means of the consistent highs and middles was tested. The obtained *t* was 2.476, also statistically significant at the 5% level of confidence, therefore confirming the third hypothesis which stated:

> Individuals predisposed toward extreme conservative social attitude will manifest greater perceptual closure reaction than those exhibiting moderate social attitude.

The closure tendencies indicated in Table 8 are graphically presented in Figure 1, where a comparatively linear trend is apparent for the authoritarian and social distance sample in contrast to the definite curvilinear relationship obtaining between perceptual closure for the consistent sample.

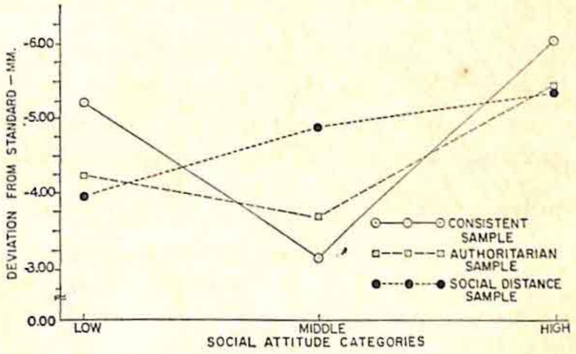


FIG. 1. Perceptual closure and the social attitude samples.

A comparison of the results of two previous exploratory studies with those of the present investigation shows striking similarities, and adds additional support to our findings. In the fall of 1952, 89 New York City secondary school *S*s were administered an F scale and perceptual closure scales similar in every important respect to the scales administered to the Houston sample. A social distance scale, however, was not employed. The only indication of social attitudes, therefore, was provided by the

TABLE 9

Comparison of the New York and Houston Studies with Regard to Perceptual Closure and Authoritarianism

Item	New York City Study		Houston Study	
	Perceptual Closure[a]	Authoritarianism	Perceptual Closure[a]	Authoritarianism
N	86	89	242	253
M	21.49	4.07	21.57	3.63
SD	3.79	.77	4.04	.82
Error	−4.51		−4.43	
% error	17.3		17.0	

[a] Standard opening: 26 mm.

F scale. One of the hypotheses of this study was: "There will be a positive relationship between a Gestalt closure scale representing perceptual closure and the California Fascism Scale representing cognitive closure" (B. Taylor 1952, p. 4). Both scales were divided at the median, and the results indicated a significant tendency for the same individuals to be consistently high or low on the scales, thus confirming the hypothesis. The middles were excluded since the median was employed for establishing the high and low categories.

Table 9 compares the F scale and the perceptual closure scale results of the New York City and the Houston studies. The number of *S*s in the former was 89, three of whom did not complete the closure scales. There were 253 *S*s in the Houston study with 242 completing this scale. Eliminees on the closure scales were 3 and 4%, respectively. Since the standard size of the configuration openings on the scales in both studies was the same (26 mm.), it can be seen that for both groups there was a highly significant tendency to underestimate the standard openings in the reproductions. The New York City group showed a 17.3% closure tendency, and the Houston group 17.0%. Both groups, therefore, exhibited practically the same degree of closure.

Comparison of the F scale means for the two groups in Table 9 indicates a difference which is significant at the 1% level of confidence (CR = 4.536). Whereas the mean of the New York City groups was approximately at the theoretical neutral point (neither agreement nor disagreement with the F scale items), the Houston group yielded a relatively low mean (disagreement with the items), which was probably due to the fact that college students tend to reject F scale items to a larger extent than do secondary school students.

When perceptual closure means were computed for each quarter of the F scale for the New York City and Houston groups, as shown in Table 10, the two groups bear a strong resemblance with regard to closure tendencies. Closure decreases with rejection of authoritarionism, although this tendency reverses itself for the low quarters. It is important to note that this reversal for

TABLE 10

Comparison of Perceptual Closure for *S*s in the Quarters of the California Fascism Scale for the New York City and Houston Studies
(Standard opening: 26 mm.)

Study	F Scale				
	Quartile Division				
	Low	Low-middle	High-middle	High	All Groups
New York City					
N	22	22	21	21	86
M	22.08	23.01	20.68	20.07	21.49
SD	4.31	3.94	3.25	3.34	3.79
Error	−3.92	−2.99	−5.32	−5.93	−4.51
% error	15.1	11.5	20.5	22.8	17.3
Houston					
N	60	62	61	59	242
M	21.77	22.12	21.72	20.62	21.57
SD	3.20	3.47	3.61	5.40	4.04
Error	−4.23	−3.88	−4.28	−5.38	−4.43
% error	16.3	14.9	16.5	20.7	17.0

lows occurs in both studies. This tendency is clearly shown in Figure 2.

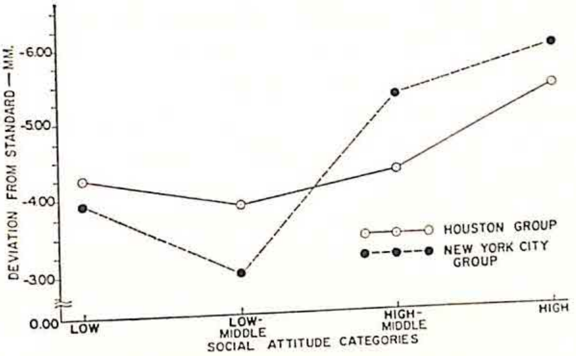


FIG. 2. Comparison of perceptual closure for the New York City and Houston groups by quarters on the California F Scale.

If the division between low and high is established at the median for the two groups, rather than at quartile divisions, the curvilinear trends are obscured, and the results support the linear interpretation as indicated in Table 11. The difference in the closure means for the lows and highs in the New York City group is statistically significant at the 1% level of confidence (CR = 2.677). The difference in means for the highs and lows in the Houston group, however, is not statistically significant. Figure 3 shows these tendencies.

TABLE 11

COMPARISON OF PERCEPTUAL CLOSURE FOR *S*s ABOVE AND BELOW THE MEDIAN OF THE CALIFORNIA FASCISM SCALE FOR THE NEW YORK CITY AND HOUSTON STUDIES
(Standard opening: 26 mm.)

Study	F Scale		
	Median Division		
	Low	High	Both Groups
New York City			
N	44	42	86
M	22.55	20.38	21.49
SD	4.16	3.23	3.79
Error	−3.46	−5.63	−4.51
% error	13.3	21.7	17.3
Houston			
N	122	120	242
M	21.95	21.18	21.57
SD	3.35	4.61	4.04
Error	−4.05	−4.82	−4.43
% error	15.6	18.5	17.0

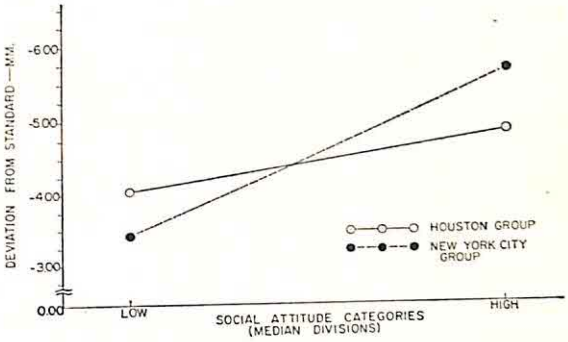


FIG. 3. Comparison of perceptual closure for the New York City and Houston groups above and below the median of the California F Scale.

A second exploratory study was designed to test the similarity of extremes hypothesis in perceptual closure. In addition to a 12 item F scale and a perceptual closure scale, 21 secondary school *S*s were administered a minority social distance scale, comparable to the one employed with the Houston group. When the consistent sample was extracted, and perceptual closure means computed for lows, middles, and highs, a curvilinear trend was obtained. Actually, the three *S*s in the consistent low category were also the three manifesting the greatest degree of perceptual closure. Since the consistent sample comprised only nine *S*s with three in each category, statistical treatment was not feasible. This exploratory study, however, suggested the desirability for more intensive investigation.

The design scale was included to test the assumption that completion occurs, and is systematically related to social attitudes, only when the figure is perceived as a single configuration. The design scale was reported as having a zero correlation with the closure scales. It was also important to examine the relationship between the design scale and the social attitude scales. These results are presented in Table 12. The re-

TABLE 12

Differences in Design Scale Means between Lows, Middles, and Highs of the Social Attitude Samples
(Standard opening: 26 mm.)

Sample	Social Attitude Category		
	Low	Middle	High
Authoritarian			
N	61	67	61
M	25.98	26.48	26.05
SD	5.02	5.21	5.22
Error	−.02	.48	.05
% error	.1	1.8	.2
Social Distance			
N	66	63	59
M	26.73	26.54	26.70
SD	5.05	5.73	4.28
Error	.73	.54	.70
% error	2.8	2.1	2.7
Consistent			
N	28	19	22
M	26.56	25.27	25.74
SD	4.67	4.19	5.43
Error	.56	−.73	−.26
% error	2.2	2.8	1.0

sults confirmed the prediction that there would be no consistent relationship between social attitudes and the size of reproduced openings of the design scale. Examination of the computed perceptual closure means for the lows, middles, and highs for the three samples indicated that all the means did not deviate from the standard (26 mm.) in either direction by more than .73 mm. None of the differences between means reached anywhere near statistical significance. The perception of a single configuration, therefore, appears to be crucial with regard to the relationship between perceptual closure and social attitudes.

The results obtained regarding the relationships between population characteristics and social attitudes also substantiate our findings thus far. Demographic characteristics were examined primarily for possible additional evidence of curvilinearity. The six population variables about which information was obtained were age, sex, religion, parental income, education, and academic major.

Table 13 shows the mean age for lows, middles, and highs for the authoritarian, social distance, and consistent samples. For all three samples, age is curvilinear, i.e., increases with intensity of attitude rather than direction. In the authoritarian sample, however, none of the differences are significant. In the social distance sample the 3.45 yrs. difference between highs and middles is significant at the 5% level of confidence (CR = 2.189), and the difference of 3.11 yrs. between lows and highs just misses the 5% level of confidence (CR = 1.903). The difference between lows and middles is not significant. None of the differences between age means for the consistent sample are significant.

In comparing the distribution of sex within the social attitude categories, shown in Table 14, a trend toward curvilinearity is also apparent; in all three samples there were more women in the extreme social attitude categories. In the authoritarian sample, only the 23.3% difference between men in

TABLE 13

Differences in Age for Lows, Middles, and Highs in the Social Attitude Samples

Sample	Social Attitude Category		
	Low	Middle	High
Authoritarian			
N	60	67	62
M	28.80	26.10	27.37
SD	12.07	7.38	8.45
Social Distance			
N	66	62	59
M	26.55	26.21	29.66
SD	9.09	8.11	9.04
Consistent			
N	28	19	22
M	28.14	27.37	31.14
SD	10.00	7.78	9.03

TABLE 14

COMPARISON OF NUMBERS AND PERCENTAGES OF MEN AND WOMEN IN THE LOWS, MIDDLES, AND HIGHS IN THE SOCIAL ATTITUDE SAMPLES

Sample	Men		Women		Men and Women	
	N	%	*N*	%	*N*	%
Authoritarian						
High	36	59.0	25	41.0	61	100.0
Middle	48	71.6	19	28.4	67	100.0
Low	29	48.3	31	51.7	60	100.0
Social Distance						
High	29	50.0	29	50.0	58	100.0
Middle	42	66.7	21	33.3	63	100.0
Low	41	62.1	25	37.9	66	100.0
Consistent						
High	12	57.1	9	42.9	21	100.0
Middle	16	84.2	3	15.8	19	100.0
Low	15	53.6	13	46.4	28	100.0

the lows and middles is significant, at the 1% level ($t = 2.728$). In the social distance sample, only the 16.7% difference between men in the highs and middles is significant, at the 5% level ($t = 2.212$). None of the differences between percentages for the consistent sample are significant, although the 30.6% difference for men in the low and middle groups just misses the 5% level of confidence ($t = 1.823$). Differences in the percentages for women, as in the case of men, are significant between authoritarian lows and middles (1% level, $t = 2.485$) and between social distance highs and middles, (5% level, $t = 2.212$).

Table 15 shows the distribution of Protestants, Catholics, and Jews among the categories of the samples. There is no discernible trend and none of the differences in any of the samples are significant.

Table 16 compares the mean estimated yearly income of the parents of lows, middles, and highs in the three samples. There is a noticeable curvilinear trend for all three samples; in each instance, a lower income mean is reported by extremes. In the authoritarian sample, however, none of the differences are significant. The trend for the social distance sample is clearly curvilinear. Lows and highs are not significantly different, but the difference between lows and middles is significant at the 5% level of confidence (CR = 2.037), and the differ-

TABLE 15

COMPARISON OF NUMBERS AND PERCENTAGES OF PROTESTANTS, CATHOLICS, AND JEWS IN THE LOWS, MIDDLES, AND HIGHS OF THE SOCIAL ATTITUDE SAMPLES

Sample	Protestant		Catholic		Jewish		All Groups	
	N	%	*N*	%	*N*	%	*N*	%
Authoritarian								
High	44	73.3	14	23.3	2	3.3	60	100.0
Middle	56	84.8	10	15.2	—	—	66	100.0
Low	48	82.8	8	13.8	2	3.4	58	100.0
Social Distance								
High	48	84.2	8	14.0	1	1.8	57	100.0
Middle	51	81.0	12	19.0	—	—	63	100.0
Low	46	73.0	15	23.8	2	3.2	63	100.0
Consistent								
High	17	80.9	3	14.3	1	4.8	21	100.0
Middle	16	84.2	3	15.8	—	—	19	100.0
Low	21	80.8	4	15.4	1	3.8	26	100.0

TABLE 16

YEARLY INCOME OF PARENTS FOR LOWS, MIDDLES, AND HIGHS IN THE SOCIAL ATTITUDE SAMPLES

Sample	Social Attitude Category		
	Low	Middle	High
Authoritarian			
N	60	65	55
M	$5660	$5720	$5200
SD	1920	1980	1860
Social Distance			
N	66	62	52
M	$5440	$6300	$5340
SD	2000	1780	1820
Consistent			
N	28	19	19
M	$5860	$7080	$4240
SD	2120	1620	1580

ence between highs and middles is significant at the same level of confidence (CR = 2.198). In the consistent sample, although the trend is curvilinear, the difference between the lows and highs is significant at the 5% level ($t = 2.209$). Also, the difference between the highs and middles is significant at the 1% level of confidence ($t = 4.075$).

The mean education level for the categories on the social attitude scales for the three samples is presented in Table 17. The trend in both the authoritarian and social distance sample is linear, although the direction of linearity is reversed. In the authoritarian sample, the difference in mean education between lows and highs is significant at better than the 1% level (CR = 2.890). The difference between education means for the lows and middles is also significant, at the 5% level of confidence (CR = 2.002). None of the differences in either the social distance or the consistent sample are significant.

The final population characteristic was academic major. This was divided into two broad categories by five staff members of the Psychology Department at the University of Houston for purposes of statistical treatment. Majors included in the first category were those described as essentially involving tangibles and were called "concrete." These included the natural sciences, biological sciences, and technologies. Majors included in the second category were those described as primarily concerned with intangibles, and were called "abstract." These included the social sciences, liberal arts, and education. The relationship between academic major judged "concrete" or "abstract" to the categories of the social attitude samples is shown in Table 18. There is a discernible tendency in all samples for the number of *S*s with "concrete" majors to increase with conservative social attitudes and, conversely, for the number of *S*s with "abstract" majors to increase with liberal social attitudes, although this tendency, using the chi square test, is not significant.

Thus far, the analysis has been confined to the authoritarian, social distance, and consistent samples with regard to the perception and population characteristics. Examination of the "inconsistent" sample also proved fruitful. This sample was defined as the inversely related quarters of the F

TABLE 17

EDUCATION LEVELS FOR LOWS, MIDDLES, AND HIGHS IN THE SOCIAL ATTITUDE SAMPLES[a]

Sample	Social Attitude Category		
	Low	Middle	High
Authoritarian			
N	59	67	59
M	3.73	3.27	3.00
SD	1.20	1.36	1.50
Social Distance			
N	65	63	58
M	3.15	3.27	3.47
SD	1.47	1.30	1.41
Consistent			
N	28	19	21
M	3.46	3.68	3.62
SD	1.24	1.03	1.45

[a] Mean education levels were interpreted as follows: Freshman—1.00; Sophomore—2.00; Junior—3.00; Senior—4.00; Graduate—5.00.

TABLE 18

Relationship between the Number of "Concrete" and "Abstract" Academic Majors to the Social Attitude Scales
(Distribution of *S*s: quartile division)

	F Scale			Minority Subscale			Consistent Quarter		
	"Concrete"	"Abstract"	Both	"Concrete"	"Abstract"	Both	"Concrete"	"Abstract"	Both
High	28	27	55	27	30	57	11	10	21
High-middle	28	30	58	25	25	50	6	8	14
Low-middle	29	34	63	29	31	60	10	16	26
Low	20	39	59	20	42	62	7	21	28
All categories	105	130	235	101	128	229	34	55	89

scale and minority subscale. Table 6 indicates that there were eight inconsistent lows, i.e., *S*s falling in the low quarter of the F scale and the high quarter of the minority subscale, 13 inconsistent low-middles, 17 inconsistent high-middles, and 13 inconsistent highs, i.e., *S*s in both the high quarter on the F scale and the low quarter in the minority subscale. The inconsistent sample comprised a total of 51 *S*s, or approximately 20% of the entire population.

The problem of inconsistency was interesting for two reasons. First, analysis of the inconsistent sample is essentially a deviant case analysis, i.e., cases which are frequently discarded as "exceptions" or "chance" variants. Second, the problem of inconsistency becomes meaningful if one assumes that the two scales measure social attitudes at two levels, the social distance scale at a relatively peripheral, overt or *explicit* level, and the F scale at a more central, covert or *implicit* level. It would follow then that the inconsistent low scorer is implicitly liberal and explicitly conservative and the inconsistent high scorer, implicitly conservative and explicitly liberal. This interpretation is weakened, however, by the fact that low scores on the F scale are probably not a reliable indication of liberal tendency, as discussed previously. The purpose of the following analysis is to compare the inconsistent high and the inconsistent low with regard to the major variables of this study.

The mean perceptual closure, age, parental income, and education for the four inconsistent categories are presented in Table 19. Although there is a tendency for inconsistent extremes to manifest less closure tendencies than the middles, none of the differences are statistically significant. Table 19 also shows that the mean age for the inconsistent highs is 23 yrs., which increases through the middle categories and reaches 33 for the inconsistent lows. The mean span of 10 yrs. between the two extremes is significant at the 5% level of confidence ($t = 2.584$). In the case of parental income a curvilinear trend is apparent, although none of the differences are significant. Education level shows a linear trend similar to age, and, as in the case of the latter, the difference between the extremes is significant, at well beyond the 1% level of confidence ($t = 4.527$). That education and age varied similarly is probably due to the correlation between the two.

Table 20 presents the numbers and percentages of *S*s in the inconsistent categories in terms of sex and religion. There was a noticeable trend for the relative percentage of men to increase with inconsistent high social attitudes and for the relative percentage of women to increase with inconsistent low social attitudes. The small num-moreover, were less reliable, as indicated by the wide variations in the reproductions of several openings in single geometric configurations. Relevant is the investigation by Goldstein and Scheerer (1941). Certain *S*s described as concrete in behavior were unable to reproduce meaningful figures from memory without changing them into more concrete ones. A line, circle, and a square, for example, were later reproduced by one *S* as windows of a church. It is possible that the presentation of relatively abstract figures to *S*s tending toward concreteness of thought produces a disrupting effect upon their reproductions, thus rendering scores unreliable. This would reduce or destroy any relationship that may have been obtained between the reproductions and other measures, and may possibly explain why reproductions of opened, abstract figures, while exhibiting completion tendencies, did not systematically relate to social attitudes.

The important finding of this study is that extremes manifested similar perceptual closure tendencies. It is necessary to emphasize, however, that this alone is not sufficient for generalizing. The basis for generalization lies mostly in the discernibly similar pattern of responses inherent in the behavior of extremes. In one sense, therefore, this study may be considered only as a clear indication of this similarity.

The interpretation of similarity of extremes is strengthened by the additional curvilinear findings. Age increased with the intensity rather than the direction of an attitude. Women tended to be more intense, in either direction, than men in their social attitudes. Parental income also showed a curvilinear tendency, with extremes reporting lower incomes than middles. Only some of these tendencies were found significant but, taken together, they support the contention of similarities.

If it is contended that extremes are similar, the next question arises as to why they are similar in the manner described. Why, for example, did not the extremes tend to open, rather than close, the openings of the configuration? If one of the related characteristics of intense social attitudes is mental rigidity, then it would seem that *S*s scoring in the extreme quarters should carefully follow instructions in reproducing figures, thus increasing rather than decreasing their accuracy. It is important to realize, however, that mental rigidity involves enduring habits instilled prior to the experimental situation. If prior life-sets involve convictions that the outline of a boy or a house ought to be complete, then this set precedes and supersedes the instructions to reproduce the figure accurately. *S*s, on the other hand, that can be characterized as less intense can respond more accurately to the instructions, having less interferences from prior sets. That extremes tend to cling to what is most familiar with the inevitable consequence of distortion is known to be characteristic of *S*s predisposed toward extreme social attitudes.

Another consideration involves clarifying the meaning of similarities and differences. The attempt in the present research has been to explore similarities at the personality level as distinct from overt social behavior. The demonstration of similarities at the personality level has not been interpreted as nullifying the social differences in the behavior of extremes. Rather, an understanding of both the similarities and the differences was considered more comprehensive than the prevalent tendency to emphasize only differences. The problem is not whether extreme liberals and extreme conservatives are similar or different, but, in what way are they similar, and in what way different? The present research has attempted to answer this question by suggesting that extremes tend to be genotypically similar and phenotypically different.

A final consideration with regard to similarities is concerned with the middle group. To the extent that the extremes can be combined at a personality level, the relatively obscured middle group also becomes focal for research in social attitudes.

One of the unanticipated findings was the positive correlation between the subscales of the social distance scale, which indicated that *S*s tended either to accept both minority and authority groups or to reject both. This was surprising, considering that included in the former were such groups as

Negroes and Mexicans, and, in the latter, Ku Klux Klan and Nazis. This finding, if substantiated by further research, indicates the importance of what may be described as *personal distance,* a general tendency to respond to all groups in a relatively similar manner, rather than the more familiar conception of *social distance,* implying a tendency to respond in widely different manners to specific groups. This tentative interpretation is supported by the finding of Hartley (1946), who, in discussing ". . . the generalized nature of tolerance" states, "Apparently there is evidence to suggest that *degree of tolerance expressed by individuals is a generalized function of the individual and is not completely determined by the specific group toward which the attitude is directed*" (p. 25).

A similar finding was also reported by Murphy and Likert (1938) in a study in which a high correlation was obtained between two randomly chosen groups for whom *S*s had previously expressed preferences. In the present study, the results support the interpretation of personal distance to an even greater extent, since the groups correlated were not random, but represented clearly opposed groups.

There is an important similarity between perceptual closure and personal distance which supports one of the major assumptions underlying this research. Extreme liberalism and extreme conservatism can be described as manifestations of the same underlying intensity which was characterized by perceptual closure. Authority and minority social distance, similarly, can be described as manifestations of the same underlying degree of acceptance which was characterized by personal distance. Both perceptual closure and personal distance, therefore, can be interpreted as genotypic constructs with opposed phenotypes.

The analysis of deviant cases—*S*s in the inversely related quarters of the F scale and minority subscale—revealed a coherent pattern of differences between the "inconsistent" low who can be described as rejecting both authoritarian beliefs and ethnic minorities and the "inconsistent" high who accepted authoritarian beliefs and ethnic minorities. The "inconsistent" lows were older, at a higher level of education, and predominately women. The "inconsistent" highs, conversely, were younger, at a lower level of education, and predominately men. With the interpretation that the F scale measures the implicit level of an attitude, and that the minority subscale measures the explicit level, there no longer appears to be a true inconsistency, but rather differences in the organization of an attitude. This tentative formulation suggests a dynamic typology involving consistent implicit and explicit and "inconsistent" implicit and explicit social attitudes. This classification would appear to be more realistic than the familiar linear ones, since it takes into consideration the important observation that individuals can exhibit various degrees of social attitudes ranging from extreme liberality to extreme conservatism depending upon, among other things, the level of organization of the attitude involved.

SUMMARY AND CONCLUSIONS

It was the purpose of this research to compare perceptual closure tendencies of extreme liberals, extreme conservatives, and intermediate groups. Extreme liberals were defined as those expressing relatively strong disagreement with the authoritarian beliefs of the F scale and accepting ethnic minorities. Extreme conservatives were defined as those expressing relatively strong agreement with the items of the F scale and rejecting ethnic minorities. It was hypothesized that *S*s exhibiting extreme social attitudes would demonstrate a similar tendency to prematurely close openings in reproducing configurations containing gaps. It was further hypothesized that either extreme would exhibit a greater degree of closure than the middle group. These hypotheses were based on observations of basic similarities discernible in the behavior patterns of extremes in social situations.

A review of the literature indicated that most studies have tended to emphasize basic differences between extremes. Examination of the California scales upon which many of these results were largely based reveals that low scores do not necessarily reflect liberality since the scale items are expressed in authoritarian terms. Many of the differences reported in the literature, therefore, may be due to this inherent weakness in the California scales.

The materials used in the present study included the F scale, a social distance scale, containing a minority and an authority subscale, two perceptual closure scales, and a design scale. Three social attitude samples were derived from these scales. The authority sample consisted of the low, middle, and high scorers of the F scale. The social distance sample consisted of the low, middle, and high scorers of the minority subscale. The consistent sample, the most important in this research, was comprised of those *S*s in the related quarters of the F scale and the minority subscale. This latter sample, although small in number, was considered to provide a more reliable estimate of social attitudes. The two closure scales correlated .71 and were combined into a single scale. The design scale, which contained gaps between several configurations, demonstrated that completion tendencies occur only when the configuration is perceived as a whole. The population consisted of 253 students attending courses at the University of Houston.

The correlation between the F scale and the minority subscale was .32. An unexpected negative relationship of −.33 between the F scale and the authority subscale indicated that *S*s accepting authoritarian beliefs rejected authoritarian groups, and *S*s rejecting authoritarian beliefs accepted authoritarian groups. An unexpected correlation of .37 between the subscales of the social distance scale indicated that *S*s tended either to accept both authority and minority groups or to reject both.

Comparison of perceptual closure means for the categories of the social attitude samples revealed a significant curvilinear relationship for the consistent sample only, thus confirming the hypotheses of the study. This result was supported by the findings of previous exploratory investigations which revealed tendencies toward curvilinearity. It was also supported by the additional findings that curvilinearity was apparent for age, which increased with attitude intensity; for sex, where a preponderance of women exhibited intense attitudes, while most men were intermediate; and for parental income, which decreased with attitude intensity. Responses on the design scale were not systematically related to social attitudes.

An analysis of the deviant cases, described as the "inconsistent" sample, included those who scored lowest on the F scale and highest on the minority subscale, and those scoring in the reverse order. The "inconsistent" lows were distinguishable from the "inconsistent" highs in age, education, and, to a certain extent, sex.

The results obtained in the present research suggested four major conclusions. First, there was the finding that extremes were similar with regard to perceptual closure. This was interpreted as supporting the contention that the structure of social attitudes of extreme liberals and extreme conservatives has basic similarities. Second, there was the finding that scores on the apparently opposed subscales of the social distance scale were positively related. This was also interpreted as indicating a basic structural similarity on the personality level. Both perceptual closure and personal distance were interpreted as genotypes which reveal themselves in apparently opposing phenotypes. Third, the finding of a coherent pattern of "inconsistent" cases with regard to the relationship between the F scale and minority subscale indicated the importance of viewing attitudes as levels of implicit and explicit organizations. Finally, the general results indicated the feasibility of utilizing the consistent sampling technique for further investigating the structure of social attitudes.

REFERENCES

Adorno, T. W., Frenkel-Brunswik, Else, Levinson, D. J., & Sanford, R. N. *The authoritarian personality.* New York: Harper, 1950.

Allport, G. W. *Personality: A psychological study.* New York: Holt, 1937.

Allport, G. W., & Kramer, B. M. Some roots of prejudice. *J. Psychol.,* 1946, **22,** 9-39.

Block, J., & Block, Jeanne. An investigation of the relationship between intolerance of ambiguity and ethnocentrism. *J. Pers.,* 1951, **19,** 303–311.

Bobbitt, J. M. An experimental study of the phenomena of closure as a threshold function. *J. exp. Psychol.,* 1942, **30,** 273-294.

Bogardus, E. A social distance scale. *Social Res.,* 1933, **17,** 265-271.

Brown, J. F. *Psychology and the social order.* New York: McGraw-Hill, 1936.

Brown, R. W. Authoritarianism and the achievement motive. *Amer. Psychologist,* 1952, **7,** 311. (Abstract)

Bruner, J. S. One kind of perception: A reply to Professor Luchins. *Psychol. Rev.,* 1951, **58,** 306-312.

Campbell, D. T., & McCandless, B. R. Ethnocentrism, xenophobia, and personality. *Hum. Relat.,* 1951, **4,** 185-192.

Christie, R., & Garcia, J. Subcultural variations in the authoritarian personality. *J. abnorm. soc. Psychol.,* 1951, **46,** 451-469.

Cohn, T. S. The relation of the F scale to a response to answer positively. *Amer. Psychologist,* 1953, **8,** 335. (Abstract)

Dombrose, L. A., & Levinson, D. J. Ideological "militancy" and "pacifism" in democratic individuals. *J. soc. Psychol.,* 1950, **32,** 101-113.

Drever, J. A dictionary of psychology. London: Penguin Books, 1953.

Eager, J., & Smith, B. M. A note on the validity of Sanford's authoritarian-equalitarian scale. *J. abnorm. soc. Psychol.,* 1952, **47,** 265-267.

Evans, R. I. Personal values as factors in anti-Semitism. *J. abnorm. soc. Psychol.,* 1952, **47,** 749-756.

Fisher, J. The memory process and certain psychosocial attitudes, with special reference to the law of Prägnanz: I. Study of nonverbal content. *J. Pers.,* 1951, **19,** 406-420.

Frenkel-Brunswik, Else. Intolerance of ambiguity as an emotional and perceptual personality variable. In J. S. Bruner & D. Krech (Eds.), *Perception and personality: A symposium.* New York: Ronald Press, 1951. Pp. 328-355.

Frenkel-Brunswik, Else. Further explorations by a contributor to *The authoritarian personality.* In R. Christie & Marie Jahoda (Eds.), *Studies in the scope and method of the authoritarian personality.* Glencoe, Ill.: Free Press, 1954. Pp. 226-275.

Gibson, J. J. The reproduction of visually perceived forms. *J. exp. Psychol.,* 1929, **12,** 1-39.

Goldstein, K., & Scheerer, M. Abstract and concrete behavior: An experimental study with special tests. *Psychol. Monogr.,* 1941, **53** (2, Whole No. 239).

Haimowitz, M. L., & Haimowitz, Natalie R. Reducing ethnic hostility through psychotherapy. *J. soc. Psychol.,* 1950, **31,** 231-241.

Hartley, E. L. *Problems in prejudice.* New York: King's Crown Press, 1946.

Hemstead, L. The perception of visual form. *Amer. J. Psychol.,* 1901, **12,** 185-192.

Hochberg, G. E., & Gleitman, H. Toward a reformulation of the perception-motivation dichotomy. In J. S. Bruner & D. Krach (Eds.) *Perception and personality: A symposium.* Durham, N. C.: Duke Univer. Press, 1950. Pp. 180-191.

Hoffer, E. *The true believer.* New York: Harper, 1951.

Hyman, H. H., & Sheatsley, P. B. The authoritarian personality: A methodological critique. In R. Christie & Marie Jahoda (Eds.), *Studies in the scope and method of* The authoritarian personality. Glencoe, Ill.: Free Press, 1954. Pp. 50-122.

Klein, G. S. The personal world through perception. In R. R. Blake & G. V. Ramsey (Eds.), *Perception: An approach to personality.* New York: Ronald Press, 1951. Pp. 328-355.

Krugman, H. The appeal of communism to American middle class intellectuals and trade unionists. *Publ. opin. Quart.,* 1952, **16,** 331-355.

Kutner, B. Patterns of mental functioning associated with prejudice in children. *Psychol. Monogr.,* 1958, **72** (7, Whole No. 460).

Luchins, A. S. Personality and prejudice: A critique. *J. soc. Psychol.,* 1950, **32,** 79-95.

Meakin, F. Mutual inhibition of memory images. *Psychometr. Monogr.,* 1903, **4.**

Meer, J. The relation between authoritarian attitudes and dreams. *Amer. Psychologist,* 1952, **7,** 312. (Abstract)

Murphy, G., & Likert, R. *Public opinion and the individual.* New York: Harper, 1938.

Roberts, A. H., Rokeach, M., & McKitrick, K. Anomie, authoritarianism, and prejudice. *Amer. Psychologist,* 1952, **7,** 311. (Abstract)

Rokeach, M. Generalized mental rigidity as a factor in ethnocentrism. *J. abnorm. soc. Psychol.,* 1948, **43,** 259-278.

Rokeach, M. Narrow mindedness and ethnocentrism. *Amer. Psychologist,* 1950, **5,** 308.

Rokeach, M. Prejudice, concreteness of thinking, and reification of thinking. *J. abnorm. soc. Psychol.,* 1951, **46,** 83-91.

ROKEACH, M. Dogmatism and opinionation on the left and on the right. *Amer. Psychologist,* 1952, **7**, 310. (Abstract)

ROSEMAN, M. Closure behavior and levels of personality integration. *J. Pers.,* 1952, **21**, 48-58.

ROSEN, E. Differences between volunteers and nonvolunteers for psychological studies. *J. appl. Psychol.,* 1951, **35**, 185-193.

SCHLESINGER, A. M., JR. *The vital center.* Boston: Houghton Mifflin, 1949.

SHILS, E. A. Authoritarianism: "Right" and "left." In R. Christie & Marie Jahoda (Eds.), *Studies in the scope and method of* The authoritarian personality. Glencoe, Ill.: Free Press, 1954, Pp. 24-49.

SROLE, L. Social dysfunction, personality, and social distance attitudes. Paper read at Amer. Sociol. Soc., Chicago, 1951.

TAYLOR, BEVERLY. An investigation of generalized closure in personality. Unpublished master's thesis, Long Island Univer., 1952.

TAYLOR, I. A. An investigation of the relationship between ethnocentrism and visual premature closure. Unpublished master's thesis, Univer. of Houston, 1951.

THURSTONE, L. L. A factorial study of perception. *Psychometr. Monogr.,* 1944, No. 4.

THURSTONE, L. L. The measurement of values. *Psychol. Rev.,* 1954, **61**, 51-52.

VETTER, G. B. The measurement of social and political attitudes and related personality factors. *J. abnorm. soc. Psychol.,* 1930, **25**, 149-189.

(Received December 6, 1958)

APPENDIX A

THE SOCIAL ATTITUDE SCALES

THE CALIFORNIA FASCISM SCALE

The following statements refer to opinions regarding a number of social issues, about which some people agree and others disagree. Please mark each statement in the left-hand margin according to your agreement or disagreement, as follows:

+1 slight support, agreement	−1 slight opposition, disagreement
+2 moderate support, agreement	−2 moderate opposition, disagreement
+3 strong support, agreement	−3 strong opposition, disagreement

______ 1. Obedience and respect for authority are the most important virtues children should learn.

______ 2. No weakness or difficulty can hold us back if we have enough will power.

______ 3. Science has its place, but there are many important things that can never possibly be understood by the human mind.

______ 4. Human nature being what it is, there will always be war and conflict.

______ 5. Every person should have complete faith in some supernatural power whose decisions he obeys without question.

______ 6. When a person has a problem or worry, it is best for him not to think about it, but to keep busy with more cheerful things.

______ 7. A person who has bad manners, habits, and breeding can hardly expect to get along with decent people.

______ 8. What the youth needs most is strict discipline, rugged determination, and the will to work and fight for family and country.

______ 9. Some people are born with an urge to jump from high places.

______ 10. Nowadays when so many different kinds of people move around and mix together so much, a person has to protect himself especially carefully against catching an infection or disease from them.

______ 11. An insult to our honor should always be punished.

______ 12. Young people sometimes get rebellious ideas, but as they grow up they ought to get over them and settle down.

______ 13. It is best to use some pre-war authorities in Germany to keep order and prevent chaos.

______ 14. What this country needs most, more than laws and political programs, is a few courageous, tireless, devoted leaders, in whom the people can put their faith.

______ 15. Sex crimes, such as rape and attacks on children, deserve more than mere imprisonment; such criminals ought to be publicly whipped, or worse.

______ 16. People can be divided into two distinct classes: the weak and the strong.

______ 17. There is hardly anything lower than a person who does not feel a great love, gratitude, and respect for his parents.

______ 18. Some day it will probably be shown that astrology can explain a lot of things.

______ 19. Nowadays more and more people are prying into matters that should remain personal and private.

______ 20. Wars and social troubles may someday be ended by an earthquake or flood that will destroy the whole world.

______ 21. Most of our social problems would be solved if we could somehow get rid of the immoral, crooked, and feebleminded people.

______ 22. The wild sex life of the old Greeks and Romans was tame compared to some of the goings-on in this country, even in places where people might least expect it.

______ 23. If people would talk less and work more, everybody would be better off.

______ 24. Most people don't realize how much our lives are controlled by plots hatched in secret places.

______ 25. Homosexuals are hardly better than criminals and ought to be severely punished.

______ 26. The business man and the manufacturer are much more important to society than the artist and the professor.

______ 27. No sane, normal, decent person could ever think of hurting a close friend or relative.

______ 28. Familiarity breeds contempt.

______ 29. Nobody ever learned anything really important except through suffering.

______ 30. America is getting so far from the true American way of life that force may be necessary to restore it.

THE SOCIAL DISTANCE SCALE

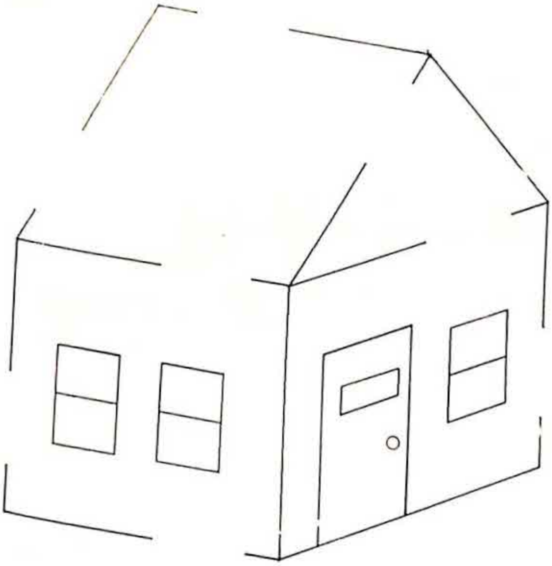
On this blank are listed the names of a number of different groups and descriptions of six different social relationships. For each group please check *all* the social relationships to which you would be willing to admit members of that group. If you would prefer not to admit members of a particular group to *any* of these six relationships, put a check mark opposite the name of that group in Column 7 (Would exclude from my country).

	1	2	3	4	5	6	7
	To close kinship by marriage	To my club as personal chums	To my street as neighbors	To employment in my occupation in my country	To citizenship in my country	As visitors to my country	Would exclude from my country
Fascists	______	______	______	______	______	______	______
Hindus	______	______	______	______	______	______	______
Puerto Ricans	______	______	______	______	______	______	______
Reactionaries	______	______	______	______	______	______	______
Mexicans	______	______	______	______	______	______	______
Nazis	______	______	______	______	______	______	______
Turks	______	______	______	______	______	______	______
Negroes	______	______	______	______	______	______	______
Ku Klux Klan	______	______	______	______	______	______	______
Totalitarians	______	______	______	______	______	______	______

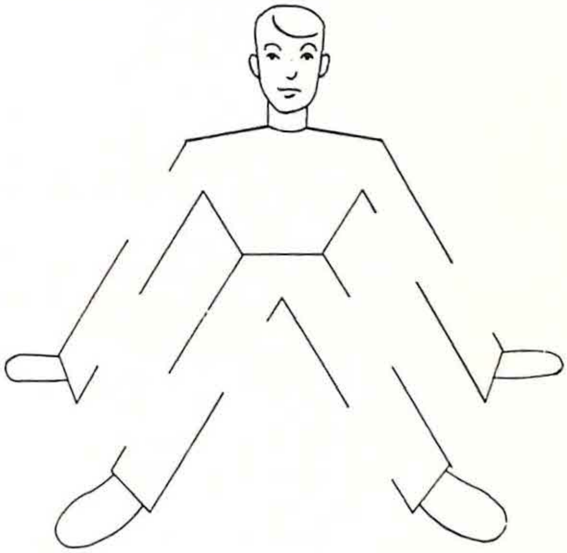
APPENDIX B

THE PERCEPTUAL CLOSURE SCALES

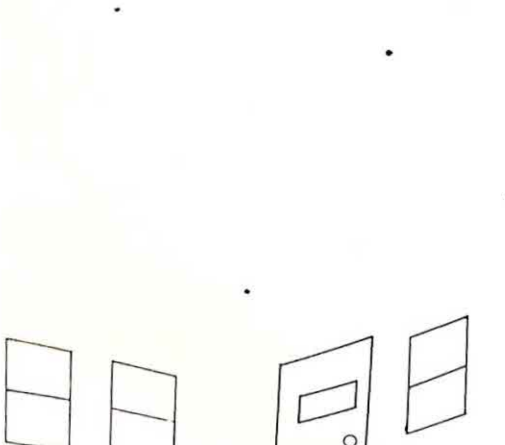
(Mean openings of original: 26 mm.)

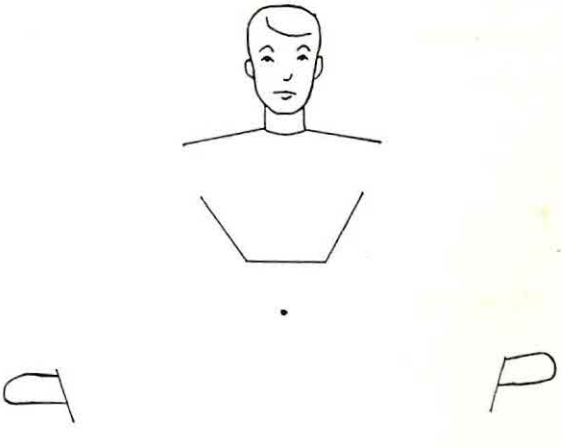


COPY THE REST OF THE FIGURE BELOW JUST AS YOU SEE IT ABOVE



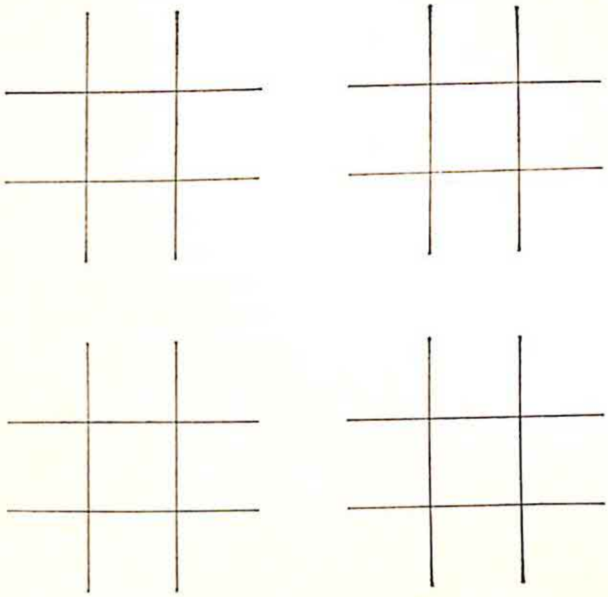
COPY THE REST OF THE FIGURE BELOW JUST AS YOU SEE IT ABOVE



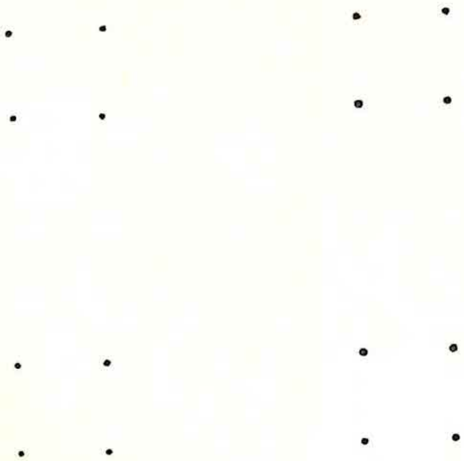


THE DESIGN SCALE

(Mean openings between configurations: 26 mm.)



COPY THE REST OF THE FIGURES BELOW
JUST AS YOU SEE THEM ABOVE



Vol. 74, No. 3 Whole No. 490, 1960

Psychological Monographs: General & Applied

PERCEPTUAL CONDITIONS OF ASSOCIATION[1,2]

SOLOMON E. ASCH, JOHN CERASO, AND WALTER HEIMER

Swarthmore College[3]

THIS is an investigation of the role of the perceptual characteristics of the stimulus in the formation of associations.

It would be natural to suppose that the dependence of association upon the properties of the stimulus constitutes a central problem, given that what is associated corresponds, often quite faithfully, to what has been perceived. For reasons that reach far into the history of thinking in our field, this has not been the case; the nature of the stimulus has largely been taken for granted. Nowhere is this more evident than in the traditional theories and experimental studies of association. Investigation carefully controls such conditions as contiguity, frequency, temporal relations, or the circumstances surrounding the appearance of the stimulus. But the properties of the stimulus itself are as a rule not systematically treated; indeed, often they are described merely by naming or pointing. We propose to show that when one takes the structure of the stimulus into account, new and neglected questions arise concerning the relation of perception to association and concerning the nature of association.

Since we rely for evidence concerning the formation of associations on the data of recall, we shall begin with the latter. Recall is the end product of a sequence of events. The distal stimulus, the proximal stimulus, the resulting perceptual process, the aftereffects it produces, and their subsequent reactivation in the form of recall, are a series of consecutive events each of which depends directly on those preceding. Many of the formulations concerning associations are statements of relation between the data of recall and these earlier events. Ideally, each of the successive steps just mentioned should find a place in the study of associations. Practice falls far short of this procedure. Before introducing the problem of this investigation, it will be helpful to see how we generally go about establishing a relation between recall and its conditions.

For illustration we choose the kind of recall termed associative. The basic observation is that the experience of *a* is followed by the recall of *b*. We consider this an instance of recall because we relate it to an earlier occasion when the corresponding stimuli, A and B, were given and observed. The correspondence between what is recalled and what was originally experienced forces us to conclude that the latter produced an aftereffect that persists and can be reactivated. It is on this ground that we distinguish between the formation and recall of associations. Further, we bring into relation the data of recall to the particular conditions prevailing at the time of original experience. Thus, if A and B were spatially contiguous, we relate the obtained recall to the condition of contiguity; similarly, if A and B appeared in a particular temporal succession, or with a given frequency.

In the instance just cited, it is usual to refer interchangeably to the preceding distal, proximal, and perceptual conditions. This ambiguity is harmless when, as is often the case, the correspondence between these respective events is so lawful as to pose no

---

[1] This investigation was supported by a grant from the National Science Foundation.

[2] For their helpfulness in making available the facilities of their Departments in the conduct of some of the experiments here reported, the writer wishes to express his gratitude to John L. Kennedy of Princeton University, Francis W. Irwin of the University of Pennsylvania, M. E. Bitterman of Bryn Mawr College, George V. Coelho and Douglas H. Heath of Haverford College, and Nelson G. Hanawalt of Rutgers University.

[3] Solomon E. Asch, now at The Institute for Advanced Study; John Ceraso, now at Yeshiva University; Walter Heimer, now at the New School for Social Research.

problem of distinction. As a rule, when A and B are contiguous, the proximal stimuli are also contiguous, and so are the phenomenal *a* and *b*. But the structural description of the stimuli is often equally informal. This practice slights the advantages inherent in an exact specification of the stimulus, advantages that are particularly important when the stimulus is perceptual.

THE PROBLEM

When we speak of the formation of an association we refer, then, to a functional connection between two experienced contents, an *a* and a *b,* corresponding to the objectively given A and B. That an association presupposes *two* terms is so obvious a requirement that it has not provoked further inquiry. We shall nevertheless see that this necessary but unexamined condition conceals ambiguities and problems of consequence about the role of perception in association.

An examination of the classical association experiment as a perceptual situation will serve to introduce the present problem. The standard procedure is to introduce the learner to a pair of stimuli, A and B, pairedness being established by means of distinctive contiguity—spatial, temporal, or spatiotemporal. Further, the stimuli are as a rule discrete and heterogeneous.

The latter property of the association experiment brings us to a surprising observation. Investigation in this region has been exclusively concerned with a particular kind of relation, namely, with stimuli each of which is *a distinct perceptual or cognitive unit.* Indeed, this property has become the criterion for the identification of an association. Investigation has confined itself to associations formed between one unit and another. The outstanding feature of this relation is that the units are (or approximate to) a mere assemblage. Each has a separate existence; take one away and little happens to the other.

This practice singles out a stimulus relation that is, to be sure, frequent in experience and important for theory. But it is neither the sole nor perhaps the representative case. There are other stimulus relations that are equally relevant to the problems in this region. Most challenging are those instances in which the respective terms belong to the same well-structured unit. Relations of the latter kind have not been investigated in connection with the problems of association, probably because the special character of the prevailing procedure was not suspected. At the same time, it seems justified to say that the various associative interpretations were intended to apply to most other instances of functional connection.

The thought from which we start may be restated as follows. The objective relations between an A and B may be of different kinds; the corresponding perceived relations may be decisive for the functional intimacy between the terms in question (and therefore for retention and recall). At the one extreme, an *a* may be so completely absorbed in a larger unit *b* that it loses psychological identity. This is the case that Gottschaldt (1926) has investigated in the region of perception; one can then no longer speak of association. At the other extreme, *a* and *b* may be discrete units; this is the case that investigation has favored under the rubric of association. There are also intermediate cases, with some of which we will be concerned, in which *a* and *b* belong to a single, tightly organized unit while retaining their respective identities.

The last-mentioned relation raises the following question. Most units consist of parts; more generally, they possess more than one property, and more than one aspect. The recall of such a unit involves the recall of its several parts or properties, as well as their order. The question of the functional intimacy between the parts of a unit is surely as pertinent as that between distinct units. There are, in short, operations necessary to the perception and recall of a unit that are formally analogous to the perception and recall of heterogeneous unit pairs.

In the light of this distinction, we ask the following questions. Are there observable differences between the association of heterogeneous perceptual units and the functional connections formed within a clearly structured perceptual unit? There is also a more fundamental question: whether the respective functional connections are qualitatively the same, or whether they involve a difference of process.

In the investigation that follows we have studied the functional connections established between stimuli related in a number of different ways. Our particular aim has been to compare the functional intimacy between heterogeneous units and between the interdependent parts of clearly structured units. The rationale of the experimental procedures follows directly from the problems just stated. We have studied the similarities and differences between stimulus conditions exemplifying different relations, but as far as possible identical in content. We have attempted in each case to specify independently the objective and the phenomenal properties of the stimulus. Since there are different particular sources of unity and nonunity, a range of conditions was studied.

The Constitutive Relation

For the first step of this investigation we chose perceptual stimuli bound by the relation we will call *constitutive*. For this purpose we turned to the following perceptual datum. A visual contour possesses at minimum two properties: it delineates a shape, but it also has a concrete material character; the contour line may be thick or thin, continuous or discontinuous, dashed or dotted. The same form, such as the circle in Figure 1 below, may thus be represented in different ways. Let us call the one property of a contour line its *form* (or F), the other its *mode* (or M).

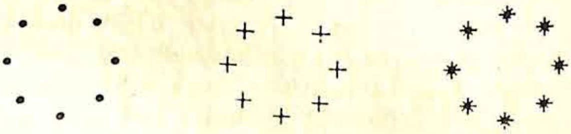


FIG. 1. Form and mode.

Concerning these properties of a visual contour we may say the following. *First,* form and mode are inseparable: a given form must appear in one or another mode. But *second,* they are not invariably united: a given form may appear in a number of modes. *Third,* the mode of a contour also has form quality: it is a form in its own right. *Fourth,* the modal form in turn may possess a mode. *Fifth,* the final modal property of a visual stimulus has a mode which is continuous. The dot and the straight line are instances of ultimate modes.

The relation between form and mode may be characterized as follows. 1. Each is a perceptual unit, the mode as much as the form, but they stand in a particular hierarchical relation. The spatial distribution of the modes generates a particular shape; thus, the shape is dependent on the modes as its constituents. The relation of dependence is such that neither term can retain its character without the other. 2. As a necessary corollary, form and modes are identical in locus; the form is where the modes are. 3. The modes generate a shape by virtue of their position, not through their particular form character. Or, they delineate a shape in terms of their point quality alone. It follows that the relation of the modal form to the encompassing form is not identical with the category of part—whole relation. 4. Finally, form and mode are clearly discriminable.

Concerning stimuli of this structure one can ask: how strongly connected do the aspects here singled out become under given conditions? When a subject (*S*) has inspected one of the figures above and attempts later to recall it, we can establish whether he has recalled one, or both, or neither of the contour properties. In short, we can determine the coherence or *jointness* of the given properties in recall, or their functional intimacy.[4] If, now, we could separate these properties and constitute them into distinct units, we would have succeeded in establishing a parallel condition, conforming to the traditional associative paradigm, to be compared with the preceding. This we propose to do in the following manner.

---

[4] The stimuli in question possess, to be sure, other properties as well—such as size, orientation, position in space—but we do not consider them in this investigation.

Procedure

The Stimulus Conditions

The stimuli selected for the first experiments are reproduced in Figure 2. Those at the left consist of different forms whose contours also differ in mode. At the right the same forms are all drawn continuously, but each is paired with a line containing the modal components of the corresponding figure at the left.

The difference between the two sets of stimuli may be described in terms of the relations they exhibit. In the first set, the relation of mode and form is constitutive. The second set of stimuli consists of the same contents, but these are now spatially separated. Separation in this case destroys the constitutive relation and abolishes the structural interdependence of the contents. These now form a pair of heterogeneous items, within which each has the same role or rank. It should be noted that form and mode were chosen to be heterogeneous, and that they were equally heterogeneous in the two series.

We can also characterize the two series in terms of their unity. By virtue of the constitutive relation, form and mode of the first series belong to the same unit; the corresponding contents of the second series are dual. The basis for this distinction is phenomenal; observers will agree that the two series differ in unity. Accordingly, we will call the first stimulus set unitary (or U) and the second set nonunitary (or non-U). This designation is intended at this point to be only descriptive.

It may be helpful to state in what respects these stimulus conditions do not differ. (*a*) Since form and mode are heterogeneous, and equally so in both conditions, the terms "unitary" and "nonunitary" refer not to the number of stimulus properties, but to the perceived relations between them. With respect to content, both sets of conditions were equally dual. (*b*) It follows that numerical singleness and duality were not the essential differences. A nonunitary pair was such not because it consisted of separated units, but rather because the units were heterogeneous. Distinct units can well function as parts of an encompassing unit; again the relation is decisive.

The U series consisted of 10 stimuli, the non-U series of 10 pairs of stimuli. Thus, form and mode correspond to the A and B terms of the traditional association experiment.[5]

The forms and modes were selected so as to be relatively simple and easily reproducible. We attempted to avoid the combination of forms and modes that might be similar or demand each other for figural reasons. (For a check on this requirement see pp. 10–11.)

The stimuli were drawn in black pencil on a white card (4 in. × 6 in.). The figures occupied, in both conditions, the center of the left-hand side of the card. The line portions of the non-U stimuli were centered on the right-hand side of the card, and were $1\frac{1}{2}$ in. long. The distance between figure and line of the latter stimuli varied from $1\frac{1}{2}$ to 2 in., depending on the shape of the former.

The contour line of a given U stimulus consisted of a sequence of identical modal forms. Six stimuli had discontinuous contours, four were drawn in continuous contour. The orientation of the modes

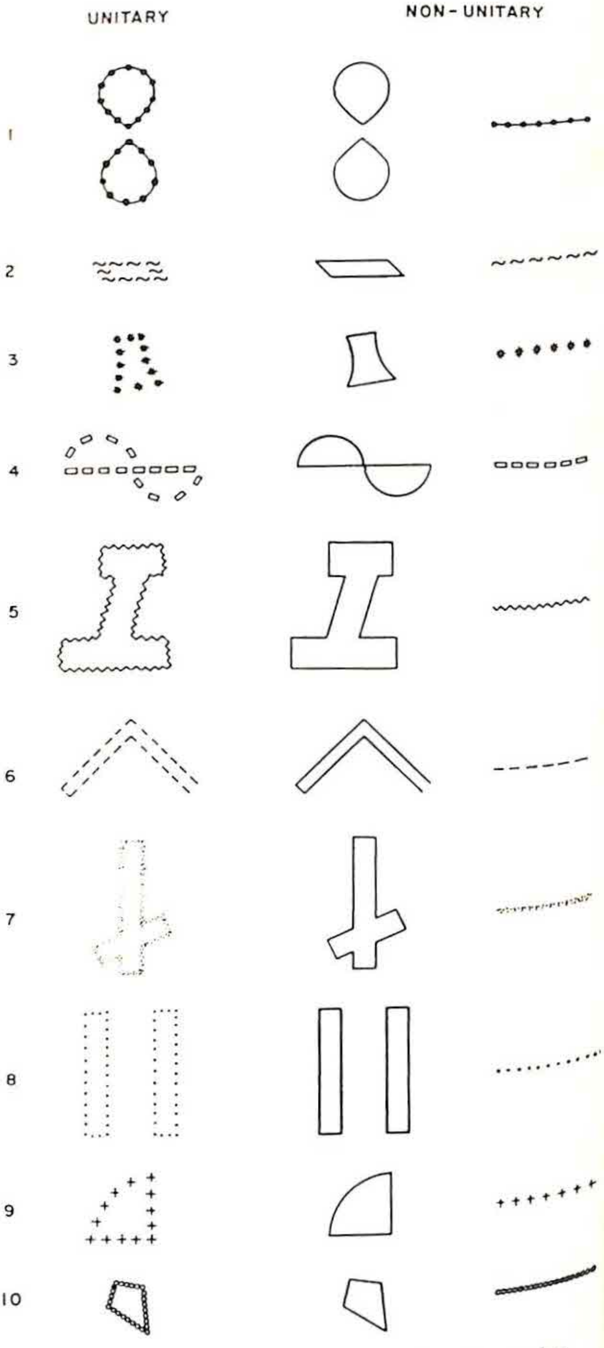


FIG. 2. Unitary and nonunitary stimulus series.

---

[5] Although the distinction between form and mode applies only to the unitary stimuli, we will, for the sake of convenience, designate similarly the corresponding parts of the nonunitary stimuli.

was kept constant, or independent of the changing orientation of the contour line, in all but two cases (stimuli 6 and 9).

The lines of the non-U stimuli were all drawn horizontally. This involved a change of orientation of the modes in the transition from the U to non-U series in three cases (stimuli 6, 8, and 9).

Conditions of Learning

Experimentation was done individually. The cards on which the stimuli appeared were placed in front of *S* against a slightly inclined wooden support stand. Each stimulus was shown singly for a period of 4 sec.; timing was done with a metronome. Each card was placed face down at the end of 4 sec., immediately exposing the next card. The stimuli were shown in the order in which they are numbered in Figure 2.

The conditions of learning were identical in each of the following variations. *S*s were instructed to inspect the figures with care so that they could later reproduce them exactly. All *S*s were thus directed to anticipate a free recall test. Particular care was taken to instruct *S*s in the non-U condition that they were to reproduce the *pairs* of stimuli on each card. It was also made clear to all *S*s that it was not necessary to recall the order in which the stimuli appeared.

The instructions were as follows:

Learning instructions: Unitary

This is a simple learning experiment. I am going to show you in succession a number of figures on the following cards.

After I have shown you the cards once, I will ask you to remember and draw the figures that you have seen. Look at each card carefully, so that you may be able to recall as many of the figures as possible. You will try to make your drawings look as much like the original as possible.

Each card will be shown once. Do not try to memorize the sequence in which the cards are shown. Any questions?

Learning instructions: Nonunitary

This is a simple learning experiment. I am going to show you in succession a number of cards.

Each card has two drawings on it—at the left is a figure, and at the right is a line drawn in a particular way.

After I have shown you the cards once, I will ask you to recall and draw the figure and the line that appeared on each card. Look at the drawings carefully so that you may be able to recall as many of the *pairs* as possible, and to draw them exactly as they appeared in the original. Please note that the task is to remember the *pairs* of drawings on each card. You will try to make your drawings look as much like the original as possible.

Each card will be shown once. Do not try to memorize the sequence in which the cards are shown. Any questions?

Conditions of Recall

Recall was tested 3 min. following the exposure of the last stimulus. The first 2 min. of this interval were occupied with a side task, consisting of the solution of anagrams. The *S*s were then interrupted and given the recall instructions.

There were three conditions of recall (with comparable groups), both in the unitary and nonunitary conditions.

Experiment 1: Free recall. *S*s were provided with blank booklets on which they reproduced all the stimuli they could recall. They were instructed to draw the contents of one stimulus card to a page, then to turn the page back, and were warned not to return to it subsequently. No strict time limit was imposed; the experiment came to an end when it became evident that *S* could make no further progress. The U *S*s were urged to reproduce either the form or mode alone when they could not recall a stimulus completely, and to guess when they were not certain. In the non-U condition they were asked to reproduce one of a pair when they could not recall its partner.

We reproduce below the recall instructions:

Free recall instructions: Unitary

Will you now draw, in this booklet, the figures exactly as they appeared earlier. Draw one figure to a page and then fold it under. You are free to recall the figures in any order.

You will have noticed that each figure was drawn with a different contour. If you remember only the figure, but not the form of the contour, just draw the figure in continuous outline. Similarly, if you remember only the form of the contour, but not the figure with which it went, draw a few samples of the contour. If you are not sure, try to guess.

Free recall instructions: Nonunitary

Will you now draw, in this booklet, the *pairs* of figures and lines exactly as they appeared earlier. Draw one pair to the page, and then fold it under. You are free to recall the pairs in any order.

If you remember one of the pairs only, draw it; if you are not sure of the other member of the pair, try to guess.

Experiment 2: Recall of mode. The booklets carried exact reproductions of the forms, all drawn continuously, one to a page. In the U condition the task was to draw the same figure with its original mode;[6] in the non-U condition the task was to reproduce the previously paired line. Except for the necessary changes the recall instructions were patterned after those reproduced above.

---

[6] *S*s were allowed to draw only a few samples of the mode when they showed impatience about going ahead.

The order of the forms in the recall test was different from the presentation order, but constant for the entire group. Again, no exact time limit was imposed; as a rule *S*s did not stay unduly long with any stimulus. Once they passed it, they were not allowed to return or make corrections. We shall refer to this as the form-to-mode (F→M) test.

Experiment 3: Recall of form. The booklets carried (on successive pages) each of the modal characters drawn as a line. U *S*s were instructed to recall and draw the figure that originally appeared in that mode; non-U *S*s were to draw the figure that had been previously paired with the given modal line. The instructions paralleled those of Experiment 2. The order of the modes was identical with that of the forms in the recall test of Experiment 2. We will call this the mode-to-form (M→F) test.

Except where otherwise stated, the *S*s were Swarthmore College students. There was an equal number of men and women in each variation.

Scoring

We will first describe the procedure of scoring the free recall reproductions. Each reproduction was first scored independently for form (F) and mode (M). The method of scoring was liberal; a form or mode was scored correct if it was recognizable as one belonging to the series. Changes in orientation, size, or proportion of the forms, and similar changes of mode, were not taken into account. The reproductions were scored by two judges; in case of disagreement they reached a joint decision. The task of scoring was relatively simple; disagreements occurred infrequently. We may note that certain highly stable and reproducible results, to be described subsequently, demonstrate that the procedure of scoring, despite an irreducible subjectivity, was quite adequate to the present purpose.

In computing an *S*'s total recall score we proceeded as follows. Each correct reproduction of a form or of a mode received a score of 1. (Thus, the complete reproduction of a U stimulus received a score of 2.) If a form or mode was reproduced several times, it was scored as correct only once. If one of the reproductions was a correct pairing, it was recorded as correct at that point; if it was not correctly paired, it was scored as correct upon first recall.

We also scored independently for accuracy of *joint* recall of form and mode, a measure that provides an index of the intimacy of connection between them. Here, we call attention to a terminological point. The joint recall of *a* and *b* in the nonunitary condition has always been called associative. Since we must leave open whether joint recall in the unitary condition involves the same process, we will employ a more neutral designation and will speak of correct or incorrect pairing. This more general term is not intended to describe the process in question; it is rather descriptive of the procedure for scoring jointness of recall. It can, of course, also be applied without prejudice to the nonunitary condition.

The structure of the task permits the following systematic and exhaustive way of scoring and ordering the data. 1. Form and mode correctly reproduced and correctly paired. We designate these recalls as +(FM), the plus sign preceding the parenthesis denoting a correct pairing. For non-U stimuli this category describes the correct reproduction of a figure and its paired line. 2. Form and mode correctly reproduced but incorrectly paired. These were instances in which a form was joined to a mode that went with another form. We refer to these reproductions as −(FM), the minus sign preceding the parenthesis denoting incorrect pairing. 3. Form alone correctly reproduced. Here were included instances in which the mode was either omitted, or an extraserial mode was introduced. Also belonging to this category were instances of a repeated reproduction of mode which had been previously scored. 4. Mode alone correctly reproduced. This category corresponded to the one just described under 3.

Scoring of the other recall tests was simpler. A reproduction (of form or mode, respectively) was first scored for correctness, whether or not it was correctly paired. It was also scored for accuracy of pairing. We thus obtained scores for the total of correct recalls and of correct pairings in Experiments 2 and 3.

Results

The Constitutive Effect

Experiment 1. The free recall data of this experiment are summarized in Table 1. Since the scores of men and women were closely similar, they will be reported jointly.

1. The total level of recall (or the recall of F and M properties regardless of accuracy of pairing) was virtually identical in the U and non-U conditions. The mean of total recalls (see column 5) was close to the 50% level, a degree of difficulty highly adequate to the present purpose. U scores ranged from 7 to 14, non-U scores ranged from 5 to 14.

2. Accuracy of joint recall was substantially more frequent in the U condition (see Columns 1 and 2). It accounted for 76.2% of all U recalls, for 49.2% of all non-U recalls. By the Mann-Whitney test the difference is significant at beyond $p = .001$. (All tests were two-tailed.)

3. Incorrect pairing was far more frequent among the nonunitary stimuli. There was a total of nine such instances in U, of 37 in non-U; (p = .001). Two-thirds of the U *S*s did not pair incorrectly, and only one *S* did so more than once. Not only did a far greater proportion of non-U *S*s mismatch the *a* and *b* terms; they also did so more often than U *S*s.

It was stated earlier that repeated reproductions of a form or mode were not counted in the scoring. This decision had consequences for the scoring of incorrect pairings involving repeated reproductions of form or mode, which will now be described.

(*a*) When an *S* paired incorrectly a form or mode which he had previously paired correctly, it was not scored as an incorrect pairing. If, for example, the *S* had reproduced F_{10} M_9, and it was preceded or followed by F_9 M_9, the latter was counted as an instance of correct pairing, and the former was scored simply as a reproduction of F_{10}, and was included in Column 3. (*b*) Occasionally, two wrong pairings occurred involving a repetition of form or mode; for example, the reproduction of F_1 M_9, and subsequently of F_3 M_9. In that case the order of recall determined the scoring; the first wrong pairing was retained and the second one was altered to read F_3, and again entered in Column 3. (*c*) If a wrong pairing, say F_4 M_8, was followed or preceded by the isolated reproduction of F_4 or M_8, the former was scored as an incorrect pairing, and the latter was dropped out of the scoring system. Thus, the absolute number of incorrect pairings was somewhat larger than appears in Table 2; however, an uncorrected procedure of scoring would not alter the finding described above. (*d*) As implied in (*a*), an accurate pairing was scored correct without exception.

4. Recall of only one of the paired terms (of F or M alone) comprised 15.4% of all U recalls, 19.5% of all non-U recalls.

5. In the case of the isolated recalls just referred to above, U stimuli produced a small but decided preponderance of form (F) recalls, and non-U stimuli showed a similar preponderance of M recalls. This finding reappears consistently in subsequent experiments (see p. 11).

The predominance of recall of modes in the non-U condition probably reflects an actual difference of complexity between them and the forms. The necessity of preparing manageable U stimuli required the selection of the simplest modes; many of these were, moreover, highly familiar conventional symbols. Why, then, were not the modes also better recalled in the U condi-

TABLE 1

Recall Scores of Unitary and Nonunitary Stimuli
Experiment 1: Free Recall Test

	N		1 Correctly Paired Recall +(FM)	2 Incorrectly Paired Recall −(FM)	3 Recall of Form Alone	4 Recall of Mode Alone	5 Mean of Total Recall
Unitary Stimuli	24	*M*	7.8	0.8	1.2	0.5	10.2
		%	76.2	7.4	11.9	4.5	
Nonunitary Stimuli	24	*M*	4.8	3.1	0.8	1.1	9.8
		%	49.2	31.4	8.1	11.4	

Note.–Each form or mode is treated as a unit of recall. Thus the figure 7.8 in Column 1 stands for the mean number of forms and modes correctly paired; the mean number of *pairings* is one half of this value, or 3.9. This is also the way to read the entries in Column 2.

The percentage values represent the number of recall units in a given category divided by the total recall. Thus, correctly paired forms and modes in the unitary series comprised 76.2% of all items recalled.

In this and in all succeeding tables, percentage values were computed by dividing the total scores of the respective categories. These values will consequently deviate slightly from, and will be more exact than, those obtained by dividing the respective means.

The total means (see Column 5) were computed by dividing the total score by *N*; these will therefore be more exact than the values obtained by adding the partial means.

tion? It is probable that form was the dominant aspect of the U stimulus (see also p. 30).

Experiments 2 and 3. These experiments employed two other methods of recall. As described earlier, we tested for recall of the mode given the continuous form (which we designate as F→M), and for recall of the form given the mode (or M→F), both under U and non-U conditions. In the recall test of Experiment 2, both groups of *S*s were given the forms drawn continuously, each appearing on the left of its page, corresponding to the position of the learning stimuli. In order to preserve the constancy of conditions between the non-U groups of Experiments 2 and 3, the learning stimuli were in the latter case drawn with the modes at the left and the continuous forms at the right. At the time of test the duplicates of the modes were also shown at the left, to correspond to their position during learning.

The results, which appear in Table 2, contain the frequencies of (*a*) correctly paired recalls and (*b*) of items correctly recalled but wrongly paired.

The results were markedly similar to those obtained with the free recall test of Experiment 1.

1. The levels of recall of U and non-U stimuli were very similar in Experiments 2 and 3. Again, mean recall was approximately at the 50% level.[7]

2. Accuracy of joint recall again was significantly higher in each of the U conditions. In Experiment 2, the percentage of correctly paired U recalls was 77.8, of non-U recalls 40.7% ($p = .001$). In Experiment 3, the corresponding percentages were 74.6 and 48.5 ($p = .01$). Differences in incorrect pairing, of course, were identical with those just described. These results are all the more notable when one considers that the test stimuli were identical with those of the learning series in the non-U condition, while differing from those of the U series.

3. The results obtained in Experiments 2 and 3 were closely similar. That is to say, F and M terms evoked each other with equal frequency.

---

[7] The scores of Table 2 refer, of course, to recall of forms *or* modes.

TABLE 2

Recall Scores of Unitary and Nonunitary Stimuli
Experiments 2 and 3: Paired Associate Recall Test

Condition	*N*		1 Correctly Paired Recall	2 Incorrectly Paired Recall	3 Mean of Total Recall
Experiment 2: Recall of Mode (F⟶M)					
Unitary Stimuli	12	*M*	4.1	1.2	5.3
		%	77.8	22.2	
Nonunitary Stimuli	12	*M*	2.0	2.9	4.9
		%	40.7	59.3	
Experiment 3: Recall of Form (M⟶F)					
Unitary Stimuli	12	*M*	3.7	1.3	4.9
		%	74.6	25.4	
Nonunitary Stimuli	12	*M*	2.7	2.8	5.5
		%	48.5	51.5	
Experiment 3a: Recall of Form (M⟶F)					
Nonunitary Stimuli	12	*M*	1.9	1.3	3.2
		%	60.5	39.5	

Note.–In this table the number of items recalled in a given category and the number of pairings in that category is, of course, identical.

This finding needs to be qualified at one point. We repeated the non-U variation of Experiment 3, this time placing the modes at the right, both during learning and recall; we refer to this condition as Experiment 3a. Thus, at the time of test the *S* had before him the right-hand member of a pair and was attempting to recall the left-hand member. The results, which will be found in Table 2, show that this procedure depressed recall significantly, to 57% of that obtained in the corresponding non-U variation of Experiment 3 ($p = .005$). Accuracy of joint recall was depressed absolutely but not relatively; the decrease did not reach statistical significance.

Since the conditions of learning were in all respects identical with those prevailing in the preceding experiments, the drop in recall of forms was a direct consequence of the method of test, and demonstrates that association of the non-U stimuli was polarized. Although the pairs were exposed simultaneously, and though the instructions did not weight the members differentially, they were psychologically successive, and were mastered in a left-right direction. There is no analogue to this asymmetry in the U condition. This result thus describes a further difference between the contrasted conditions: U-pairings were not polarized, in contrast to non-U pairings.

4. Experiments 2 and 3 did not produce higher levels of correct pairing than did Experiment 1. This result is somewhat surprising when one considers that recall was of the paired-associate form, in contrast to the free recall procedure of Experiment 1. It would have been reasonable to expect that some of the isolated recalls of Experiment 1 would be converted into correct pairings in the present variations. A hint of what occurred is provided by the increased frequency of incorrect pairings, both in the U and non-U variations. Apparently, aided recall introduced its own constraints; a *S* had to respond correctly at a given point, or be penalized, whereas the free recall procedure permitted him to repeat any combination of form and mode, and thus to correct his earlier responses. That the levels of total recall were closely similar to those of Experiment 1 is not surprising, since the conditions of learning were not altered.

Confirmatory Experiments

Experiment 4: Combined U and non-U Conditions. The preceding experiments compared homogeneous U and non-U series. The following variation combines these within the same experiment, a procedure that permits each *S* to serve as his own control.

From the respective series of Experiment 1 we derived two new series, each consisting of 5 unitary and 5 dual stimuli. There were two groups of *S*s. The stimuli that were given in U form to one group appeared in non-U form for the other group, and conversely. In assigning stimuli to the two categories, we chose sets equated for difficulty on the basis of the data of Experiment 1. (The U stimuli for group 1 were: 1, 4, 6, 7, and 10.)

We also rotated the order of presentation of the stimuli, so that each stimulus would appear in each of the 10 positions. The order of presentation of U and non-U stimuli was, for the first *S*, ABBABAABBA. This order was shifted one step for each succeeding *S*.

The instructions described in detail the two kinds of stimuli, with the help of an illustrative example of each. Again, the *S*s were set for a test of free recall. They were cautioned explicitly to "make every effort to pay equally close attention to each card."

Testing was with the method of free recall. *S*s were urged to reproduce part of each stimulus when they could not recall it entirely and to guess when not certain. The procedure was in all respects identical with that of Experiment 1.

There were 20 *S*s, 10 in each group, equally divided among men and women.

a. Results. The overall comparisons between the U and non-U stimuli, obtained by pooling the results of both groups, appear in Table 3.

1. The difficulty of the tasks, measured in terms of recall of content, was closely similar to those of Experiment 1.

2. U and non-U stimuli produced virtually identical total recall in one series; in the other series the U stimuli showed higher recall.

3. Under the present conditions, the differences between U and non-U stimuli in accuracy of joint recall were considerably accentuated. Correctly paired recalls accounted for 88.7% of all U recalls, for 29.3% of all non-U recalls ($p = .001$).

4. The differences in inaccurate pairing were even sharper. There was only one such instance in U, but inaccurate pairings comprised 37.8% of all non-U recalls.

5. Sixteen out of 20 *S*s were superior in pairing correctly the U stimuli; three *S*s performed equivalently; and there was one difference in favor of the non-U stimuli. (Of the *S*s who went contrary to the predominant trend, three recalled very poorly throughout, and one obtained a nearly perfect recall.)

6. Of the total of 10 stimuli, 9 produced more correct pairings in U; one stimulus failed to show a difference.

7. The reproductions showed a remarkable lack of confusion between the two sets of stimuli. (*a*) There was a total of 16 instances of incorrect pairing. Only one of these involved the combining of a form from one set with a mode from the other set; all other mispairings occurred within the same set. (*b*) Further, the U or non-U character of the correctly paired stimuli was almost faultlessly retained. In only one instance were the form and mode of a U stimulus recalled as separated; in no case was a pair of non-U stimuli reproduced in a joined manner. These results lend support to the conclusion that the boundaries between the two sets of stimuli were strong.

Under the present conditions, the differences between the stimulus conditions increased sharply, when compared with Experiment 1. It would appear that, in the competition for recall, U stimuli had a decisive advantage. This finding might be questioned if the *S*s had deliberately favored the U stimuli during learning. As mentioned earlier, the instructions attempted to guard against this contingency. As a further check, the *S*s were questioned about their learning procedure at the conclusion of the experiment. Twelve *S*s reported paying equal attention to both kinds of stimuli, 5 *S*s favored the U, and 3 reported that they had studied the non-U stimuli more carefully. According to these reports, there was no particular bias in the procedures the *S*s followed.

It may be pertinent to mention that the greater number of *S*s expressed the view that U stimuli were easier because simpler. "There was only one thing to remember in the single figures, two things to remember in the double figures"; "The double ones sort of split my attention"; "There was more to remember with the double figures—two separate objects."

The results thus confirm the main findings previously reported concerning the differences between the U and non-U conditions.

Experiment 5: Randomized Pairing of Form and Mode. In the preparation of stimuli for the preceding experiments we

TABLE 3

RECALL SCORES OF UNITARY AND NONUNITARY STIMULI

Experiment 4: Combined Unitary and Nonunitary Condition — Free Recall Test

	N		1 Correctly Paired Recall +(FM)	2 Incorrectly Paired Recall −(FM)	3 Recall of Form Alone	4 Recall of Mode Alone	5 Mean of Total Recall
Unitary Stimuli	20	*M*	4.7	0.1	0.3	0.3	5.3
		%	88.7	0.9	5.7	4.7	
Nonunitary Stimuli	20	*M*	1.2	1.6	0.5	0.8	4.1
		%	29.3	37.8	12.2	20.7	

sought to avoid combinations of forms and modes that might be similar or outstandingly fitting. Since there was no assurance that we excluded such selectiveness completely, the following variation was done as a check upon this possibility.

The forms and modes of the preceding experiments were now randomly paired. We numbered the modes from 1 to 10, drew these numbers and assigned each to the forms, in the order they had in the earlier learning series. We thus obtained a new pairing of F and M properties. Three of the original stimuli (Nos. 5, 7, and 8) were line forms; these we changed into closed areal forms (in preparation for a further variation to be described in a subsequent section). These stimuli, which will be employed frequently in subsequent experiments, are reproduced in Figure 3.

The procedure, instructions, and methods of scoring were identical with those of Experiment 1. The test was one of free recall. There were 24 *S*s in each variation, 12 men and 12 women. The summary data appear in Table 4.

1. Recall of content, or of the forms and modes, was again virtually identical in the U and non-U conditions. The mean levels of recall continued to be similar to those of Experiment 1 (see Table 1).

2. U stimuli produced far more correct pairings than non-U stimuli (65.5% and 38.1%, respectively, of all recalls: p = .001). Inaccurate U pairings were far less frequent than inaccurate non-U pairings (p = .001). As in Experiment 1, U *S*s either did not mismatch, or did so no more than once; mismatching occurred among more non-U *S*s, and with higher frequency.

3. Also in agreement with the results of Experiment 1 is the finding that isolated recalls of U forms were more frequent than of U modes, but that in the non-U series there was a predominance of recall of modes.

We have thus confirmed the previous findings, and demonstrated that the effect persists under modified stimulus conditions.

Index of the Constitutive Effect

Since the principal difference between U and non-U stimuli concerns accuracy of joint recall, it would be helpful to express the effect in such a manner as to permit comparison among different experimental conditions. One index that recommends itself for this purpose is the ratio of correct to incorrect pairings. In order to avoid complications due to zero scores, we adopted as the index the difference between correct and incorrect pairings divided by their sum. This index varies between ±1. When incorrect pairings are zero, the value is 1; when correct pairings are zero the value is

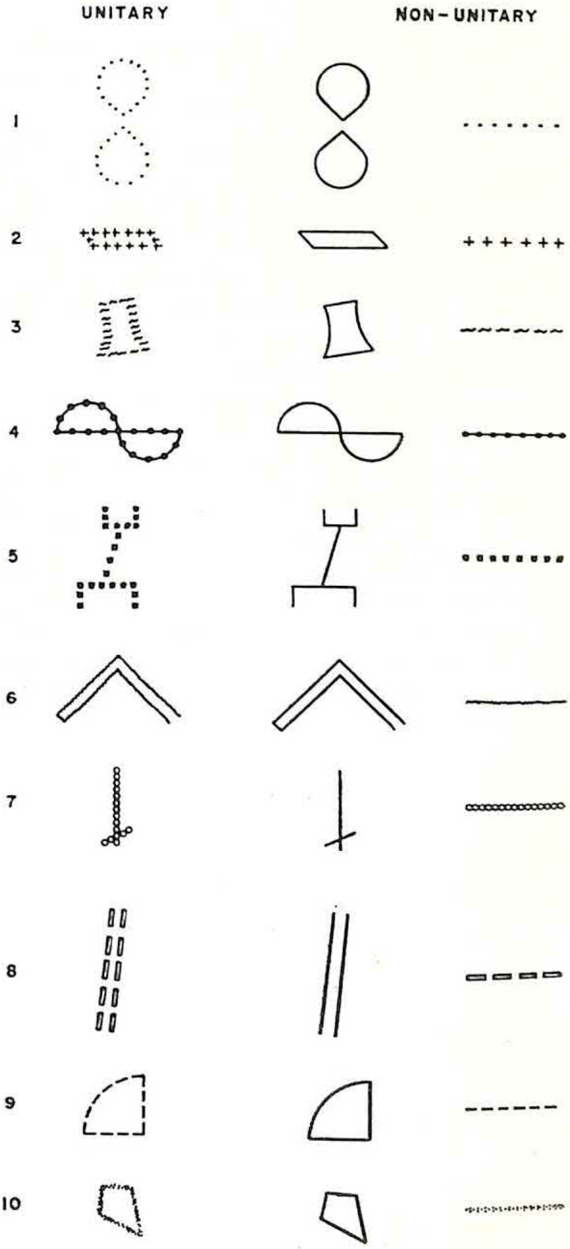


FIG. 3. Unitary and nonunitary stimulus series: Experiment 5.

TABLE 4

Recall Scores of Unitary and Nonunitary Stimuli
Experiment 5: Free Recall Test

	N		1 Correctly Paired Recall +(FM)	2 Incorrectly Paired Recall −(FM)	3 Recall of Form Alone	4 Recall of Mode Alone	5 Mean of Total Recall
Unitary Stimuli	24	*M*	6.3	1.3	1.6	0.4	9.7
		%	65.5	13.8	16.8	3.9	
Nonunitary Stimuli	24	*M*	3.7	3.3	1.0	1.8	9.6
		%	38.1	33.8	9.9	18.2	

−1; and the index takes on the value zero when both kinds of pairing occur with equal frequency.

We accordingly computed the values of the index for each *S*. Table 5 summarized the data for Experiments 1 to 5, and of a number of subsequent experiments employing the method of free recall. We computed the median values of the index in each condition, the proportion of *S*s with exclusively correct pairings, and the proportion of *S*s with predominantly incorrect pairings. In obtaining these values we grouped together Experiments 1 and 5, and also Experiments 2 and 3.

The U and non-U series differed consistently for all selected values of the index. The U medians of Experiments 1 to 5 ranged from 1.00 to 0.7, the non-U medians from 0 to 0.3. The proportions of +1 values

TABLE 5

Index of Pairing Accuracy

$$\left(\frac{\text{Correct Pairings} - \text{Incorrect Pairings}}{\text{Correct Pairings} + \text{Incorrect Pairings}}\right)$$

Conditions	*N*	*Md.*	+1 Scores (in %)	Negative Scores (in %)
Experiments 1 and 5 combined:				
Unitary Stimuli	48	1.0	54	2
Nonunitary Stimuli	48	0.3	21	23
Experiments 2 and 3 combined:				
Unitary Stimuli	24	0.7	42	13
Nonunitary Stimuli	24	0.0	0	25
Experiment 4				
Unitary Stimuli[a]	20	1.0	95	0
Nonunitary Stimuli[a]	20	0.0	25	45
Experiment 7				
Unitary Stimuli	16	0.6	94	0
Nonunitary Stimuli	16	0.1	50	25
Experiment 11	24	0.0	25	21
Experiment 12				
"Coherent" Stimuli[a]	24	0.9	50	4
"Not coherent" Stimuli[a]	24	0.1	21	33

[a] Based on 5 stimuli.

and of negative values exhibited pronounced differences in the same direction.

The following analysis provides further evidence of the differences in pairing between and within units. When the method of test was that of free recall, there were instances of the isolated recall of both the form and mode of the same stimulus, or of the mispairing of either with another member of the series. The procedure of testing leaves the *S* free to pair correctly at some later point those items which he had previously recalled but not paired, or which he had mispaired. We therefore ask: When the form *and* mode of a given stimulus were recalled separately, what was the likelihood that they would also be eventually correctly paired?

Table 6, which contains the relevant data for Experiments 1 and 5, demonstrates a marked and consistent difference between U and non-U stimuli in this respect. There was a total of 10 instances in the U variation of isolated recall of form and mode of a given stimulus without eventual correct pairing of the same terms; in the non-U variations there were 42 such instances. That is to say, when both form and mode of a unitary stimulus were recalled, they were as a rule also correctly reunited; this was far less often the case for nonunitary stimuli.

Control Experiments: Recognizability and Intraserial Interference

We have inferred that the preceding results were directly a function of the respective stimulus relations—of the constitutive relation and of the relation of duality between paired heterogeneous contents. Before one can safely draw this conclusion, it is necessary to determine whether the stimulus conditions differed also in other respects, and whether the latter might be responsible for the obtained results. The present experiment and the one following are concerned with this question.

Experiment 6: Recognizability of Unitary and Nonunitary Stimuli. It is by no means established that the U and non-U stimuli were equally distinct or identifiable. The rationale of this investigation obviously requires that the experimental effect should not be a function of differences of this kind. The object of this variation, accordingly, was to compare the identifiability of the forms and modes in the two series.

a. Procedure. As the criterion of distinctness or clarity we chose the recognizability of the forms and modes. Except for the final test, the stimuli, instructions, and other conditions of learning were identical with those of Experiment 5. As before, the *S*s were instructed to pair form and mode in anticipation for a test of free recall. However, the test that followed, and which will now be described, was one of recognition.

The test booklets carried exact reproductions of each form and mode, one to a page; each appeared in the center of the page. The forms were all drawn continuously, and the modes were in linear array. Included in the test were 10 new forms and 10 new modes. The task was to check a "yes" or "no" printed on each page, indicating whether or not the given stimulus had been part of the learning series. The *S*s were informed that new stimuli would be interspersed in the test series. Again, the test of recognition came 3 min. following the presentation of the learning series.

The sequence of stimuli was as follows. The old forms and modes were each divided into 2 blocks of 5. To each of these 4 sets we added a set of 5 new forms or modes. Thus, the sequence consisted of 4 blocks of 10 stimuli, each containing 5 old and 5 new forms or modes. The order of the blocks was varied with succeeding *S*s; there was thus a total of 4 orders. Also, the order of the critical stimuli within each block was rotated from *S* to *S*; the noncritical stimuli retained the same position within a block.

The *S*s were Princeton University undergraduates, 16 in each group.

b. Results. The percentages of correct recognitions of the critical stimuli, and of

TABLE 6

FREQUENCY OF FAILURES TO PAIR RECALLED FORM AND MODE BELONGING TO THE SAME STIMULUS

	Unitary Stimuli		Nonunitary Stimuli	
	F	%	*F*	%
Experiment 1	4	4.1	16	21.6
Experiment 5	6	7.3	26	37
Experiment 7	53	38	84	70

TABLE 7

RECOGNITION OF FORMS AND MODES
(in %)
Experiment 6

	Critical Stimuli	Noncritical Stimuli
Unitary Stimuli		
Form	86	91
Mode	69	67
Nonunitary Stimuli		
Form	90	94
Mode	83	92

correct nonrecognitions of the noncritical stimuli, appear in Table 7.

(*a*) The level of recognition of the critical stimuli was, with one exception to be noted below, quite high. (*b*) Recognition of critical forms was superior to recognition of critical modes. By the sign test, this difference was significant in the U series ($p <$ .01), not in the non-U series. (*c*) Critical forms were equally recognizable, whether they were unitary or not. (*d*) However, the recognition of non-U modes was higher than that of U modes ($p = .02$). (*e*) Nonrecognition of the noncritical stimuli was, with one exception, slightly but consistently superior to the recognition of critical stimuli.

The test stimuli were most similar to those of the non-U series. This circumstance fails, however, to account for the difference in recognition of unitary and nonunitary modes, when one considers that the forms were equally recognizable in the two series. We conclude that there was a tendency for form to be dominant over mode in the U series. This is not surprising in the light of the distinctive functions of form and mode in the U stimulus, and of their comparative similarity of function in the non-U stimulus.[8]

While it is difficult to state precisely the relation between these data of recognition and the earlier data of recall, we may draw the following limited conclusion. Since the level of recognition sets limits to the pairings that can occur, differences of stimulus clarity did not favor the U series. If anything, the reverse was more likely the case; to the extent that this factor played a role it could only favor pairing in the non-U series.[9]

As a further demonstration of the same point, we may cite the following analysis. From the recognition data we obtained a measure of "synthetic pairing," namely, of the incidence of recognition of form and mode belonging to the same unit or to the same pair. This measure presumably sets a limit to the pairings that can occur. There were 95 such synthetic pairings in the U-group, and 122 in the non-U group ($p =$ .02). These comprised 76% of all recognitions in the U-series, and 90% of all recognitions in the non-U series. Again, we have to conclude that the conditions of registration tended to favor the non-U series; if so, the recall data of the preceding experiments tend to underestimate the effect of the constitutive relation on pairing.

Experiment 7: The Role of Intraserial Interference. It might be claimed that the stimulus series were not equally homogeneous. The non-U series consisted of forms all of which had continuous boundary lines, and of other forms all of which appeared as line aggregates. This circumstance might make them more vulnerable to intraserial interference than the apparently more individualized U stimuli.

In the variation to be now reported, the *S*s were afforded an opportunity to become familiar with the individual forms and modes prior to the main task. The procedure of familiarization raised notably the distinctness and recallability of the terms before the main task was introduced. This step could only have the effect of reducing the levels of subsequent intraserial interference, and of equalizing the levels in the respective series. We reasoned that the experimental effect should be similarly reduced if it were a function of serial interference.

---

[8] Consistent with the present data is the finding reported earlier of a slight but regular superiority of form over mode recall in the U series. The results are not, however, consistent with the corresponding finding that non-U modes were somewhat better recalled than forms.

[9] One must caution, however, against inferring too directly from recognition to recall. When the stimuli (of each series) were ranked, respectively, for ease of recognition in the present variation and for ease of pairing (in Experiment 5), the correlations were low and lacked significance.

a. Procedure. The following was the procedure of familiarization. The forms and modes of Experiment 5 were drawn on separate cards, and in the positions they previously occupied—the forms at the left, the modes at the right.

One half of the *S*s were shown first the set of 10 forms, in succession. The period of exposure of each form was 4 sec. The series was shown 3 times; before each trial the order was changed by shuffling the series. *S*s were instructed to study the forms so as to be able to reproduce them correctly. The third trial was followed by a test of free recall. Thereupon, the *S*s were shown the series of 10 modes under identical conditions, and again recall was tested. The other half of the *S*s were first shown the modes, and then the forms.

After a lapse of 3 min. we proceeded to a repetition of Experiment 5. *S*s were assigned to the unitary or nonunitary groups on the basis of the preceding tests; the two groups were closely matched. The procedure of Experiment 5 was maintained without change, except that it was possible to point out during the instructions (with the help of an illustrative stimulus) the relevance of the preceding familiarization to the task.

b. Results. The procedure of familiarization produced a high level of recall of forms and modes (Table 8, first column). Further, the level of recall was virtually identical for the two groups of *S*s. There was a slight but consistent superiority in the recall of modes; for the entire group of 32 *S*s, the superiority of modes was, by the sign test, significant at beyond .01.

The data concerning pairing are also summarized in Table 8.

1. The chief finding is the superiority of joint recall in the U condition. *S*s in the latter group showed a mean of 5.3 correct pairings; the mean of non-U pairings was 2.3 ($p < .001$). Each unitary stimulus produced more correct pairings than its nonunitary counterpart.

2. Under the present conditions the U and non-U series did not produce significantly different frequencies of incorrect pairings; the respective means were 1.4 and 1.9.

Familiarization significantly raised the frequency of correct U pairings, compared with the level obtained in Experiment 5 ($p < .01$), but did not significantly alter the level of non-U pairings. Further analysis demonstrates, however, that familiarization did not alter the *relative* frequencies of accurate and inaccurate pairings in either series, in comparison with Experiment 5. In order to allow for the differences in absolute level of recall, we took the total of accurate and inaccurate pairings, and computed the proportion of each to the total. The results appear in Table 9. The ratios are substantially similar to those of Experiment 5. The differences in absolute level tell us that familiarization increased the at-

TABLE 8

Mean Recall Scores of Unitary and Nonunitary Stimuli Following Familiarization
Experiment 7: Free Recall Test

	N	1 Mean of Total Recall before Pairing	2 Correctly Paired Recall +(FM)	3 Incorrectly Paired Recall −(FM)	4 Recall of Form Alone	5 Recall of Mode Alone	6 Mean of Total Recall after Pairing
Unitary Stimuli	16	18.0	5.3	1.4	2.4	2.6	18.3
Nonunitary Stimuli	16	17.8	2.3	1.9	4.1	4.8	17.1

Note.–The mean values in Column 1 are the recall scores obtained following familiarization and before pairing. The values f Column 6 were obtained during the final test of free recall.

TABLE 9

PROPORTION OF CORRECT PAIRINGS TO TOTAL PAIRINGS
Experiments 7 and 5

	Correctly Paired Recall +(FM)	Incorrectly Paired Recall −(FM)
Experiment 7		
Unitary Stimuli	78	22
Nonunitary Stimuli	54	46
Experiment 5		
Unitary Stimuli	83	17
Nonunitary Stimuli	53	47

tempts at pairings, but not the proportions of the kinds of pairing.

3. It follows that isolated recalls must be more frequent in the non-U series. This was the case; the mean of isolated non-U recalls was 8.9, of isolated U recalls, 4.9.

4. The index of pairing accuracy revealed a clear-cut difference between the two experimental conditions (see Table 5).

5. Finally, we again found that when form *and* mode of the same stimulus (or pair) were recalled, they were also correctly paired with far higher frequency when unitary. Of all recalled forms and modes belonging to the same stimulus (or pair), 62% were correctly paired by the U group; this value dropped to 30% in the non-U group (Table 6).[10]

6. The total mean recall, regardless of pairing, was 18.3 for the unitary series, and 17.1 for the nonunitary series; this difference was not significant (see Table 8, last column).

These results were obtained after a high degree of familiarization with the contents of the series. There can be little doubt that this procedure, which equalized the identifiability of the stimuli for the two groups, also reduced and tended to equalize interference during subsequent pairing. Had the experimental effect been a function of interference, it should have been markedly reduced under these conditions. Actually, the effect became more pronounced. The ratio of correct U pairings to correct non-U pairings was 2.3; in the corresponding variation of Experiment 5, the ratio was 1.7.

We conclude that the experimental effect was not a function of differences of intraserial interference. Further, it seems to us that this variation, which eliminated the confounding due to the concurrent tasks of mastering the terms and pairing them, provides a more accurate approximation of the experimental effect, which the earlier results underestimated.

These findings also confirm the conclusions of the preceding variation. Equalizing the distinctness or identifiability of the stimuli had the effect of accentuating the difference between the conditions.[11]

The Scope of the Unitary Effect

In this section we continue the investigation by varying further the conditions of learning and recall, in order to establish the scope of the unitary effect.

The results so far reported were based entirely on the data of recall. In memory investigation an obvious and important distinction is that between the presence or retention of an association (or, more generally, a pairing) and its recall, since a functional connection may be present but not available. We can only conclude from what has preceded that unity was a condition favoring recall; it remains to be shown whether the effect extended to the formation proper of pairs.

---

[10] Table 6 shows that failures to pair were frequent in this condition, in comparison with earlier variations. This was directly a consequence of the procedure of familiarization.

[11] Why did familiarization raise the absolute level of correct U pairings, and not of non-U pairings (compared with the levels of Experiment 5)? We interpret these results in the light of Experiment 6. There we saw that the recognizability, and presumably the perceptibility, of the U modes was comparatively low. Familiarization increased the distinctness of U modes, and thus raised joint recall. But the non-U series, being more perceptible, could profit less from the procedure of familiarization.

As a first step in this direction we employed the test of *matching*. This procedure reduces considerably the burden placed upon the recall of contents. When the *A* and *B* terms are given at the time of test, the sole task of the *S*, once he has recognized them, is that of recalling the pairing they had initially. The procedure of matching thus provides a more direct test of the recall of an association, or of pairing.

Experiment 8: The Procedure of Matching—Intentional Learning. Experiment 5 was repeated with new groups of *S*s. The experiments were now done in groups ranging in size from 6 to 10.

The stimuli were projected on a screen with standard size lantern slides. The projected size of the images was controlled, by setting the width of stimulus 4 at 12 in. *S*s sat at a distance of 6 to 12 ft. from the screen. They were placed as close to the center as possible, their distance from the center of the screen not exceeding 35° — 40°. (As a further precaution, the *S*s were asked, at the conclusion of the experiment, to state in writing whether they had been able to see the stimuli distinctly; all answered in the affirmative.)

The other conditions of learning were unchanged. Learning was intentional; as before, the instructions set the *S*s to anticipate a test of free recall. An illustrative card was now introduced during the instructions, containing a unitary (or nonunitary) stimulus, in order to insure that the task would be understood (and also to make possible a comparison with a subsequent variation to be described below, which necessitated the illustrative stimulus). Also, an interval of 4 sec. was introduced between the exposures of successive stimuli, (again for the purpose of comparison with a subsequent variation that required this procedure).

A side task of anagram problems immediately followed the presentation of stimuli. After 1½ min., booklets containing the forms, drawn in continuous outline, were distributed; each form appeared on a separate page. The order of the forms was varied from *S* to *S*. A matching sheet containing the 10 modes was distributed at the same time, each drawn as a straight line. The modes were placed in a random sequence (constant for all *S*s), and were numbered in order from 1 to 10. The instructions for the matching test were then read. The *S*s were directed to turn to the first form in the booklet, to select from the matching sheet the mode that went with it previously, and to enter its number next to the form. They were to guess when uncertain, and they were free to select a given mode more than once, but they were not to return to a page once they had passed it.

When this test was completed, the matching sheets were collected. Thereupon, the *S*s were given a modified test of recognition. They were asked to inspect again the forms in their booklet, and to indicate whether they recognized each as one that had appeared in the original series by writing "yes" or "no" next to it. *S*s were warned not to alter their earlier matching responses; this they were in any case unlikely to do, since they were at this point without the matching sheets.

The *S*s were women students at Bryn Mawr College, 20 in the U variation and 19 in the non-U variation.

TABLE 10

RECALL SCORES OF UNITARY AND NONUNITARY STIMULI

Experiments 8 and 9: Matching Test

	Unitary stimuli		Nonunitary stimuli	
	N	Mean Correct Matchings	*N*	Mean Correct Matchings
Experiment 8: Intentional	20	6.1	19	3.2
Experiment 9: Incidental	11	4.1	15	1.7

Unitary stimuli produced clearly superior matchings, as Table 10 shows. The mean of correct U matchings was 6.1; the mean of non-U matchings was 3.2 ($p = .0002$).

It would have been instructive to compare the levels of matching with the levels of correctly paired recall in the earlier experiments. This we are unable to do, since the present procedure involved a change in the mode of exposure, and the introduction of an interval between exposures. The means of correct matchings are virtually identical with the means of correctly paired recalls in Experiment 5, but for the reasons stated we can draw no inference from this finding.

The data concerning the recognition of forms, and their relation to accuracy of matching, appear in Table 11. Failures to recognize the forms were infrequent, almost equally so in both variations, the proportions of such occurrences being 7.5% in U, and 10.0% in non-U.[12] Although the procedure of testing suffered from certain shortcomings—it was preceded by a test of recall, and it lacked noncritical items—the levels of recognition were closely similar to those obtained in Experiment 6 (see Table 7). We did not test for recognition of modes, but it is reasonable to assume that these levels were again similar to those of Experiment 6. The results just described highlight the fact that the U and non-U conditions differed markedly in matching at the same time that they were virtually identical with respect to recognition of forms.

---

[12] Despite the small number of failures to recognize, they were not randomly distributed among the stimuli; certain forms (e.g., 1,5,6) were always recognized, by both groups; others (e.g., 9,10) produced relatively high failures to recognize in both groups.

When unitary forms were recognized, they produced correct matches in 63.8% of the cases; correctly recognized nonunitary forms produced correct matches in only 31.6% of the cases. Thus, recognition of U forms went with correct matching in the majority of cases, and recognition of nonunitary forms went with incorrect matching in the majority of cases.[13]

The present results demonstrate that the superiority of the unitary condition persists when the burden of item recall is reduced or eliminated.

All of the preceding experiments were done under conditions of intentional learning. It would be possible to maintain that the differences obtained between unitary and nonunitary stimuli were a function of learned ways of handling the respective tasks. If so, a procedure of incidental learn-

TABLE 11

Relation of Recognition to Accuracy of Matching
Experiments 8 and 9

		Experiment 8: Intentional Learning					
		Unitary Stimuli ($N = 20$)			Nonunitary Stimuli ($N = 19$)		
		Correct Matching	Incorrect Matching	Total	Correct Matching	Incorrect Matching	Total
Recognition	*F* %	118 63.8	67 36.2	185	54 31.6	117 68.4	171
Failure of Recognition	*F* %	3 20.0	12 80.0	15	6 31.6	13 68.4	19

		Experiment 9: Incidental Learning					
		Unitary Stimuli ($N = 11$)			Nonunitary Stimuli ($N = 15$)		
		Correct Matching	Incorrect Matching	Total	Correct Matching	Incorrect Matching	Total
Recognition	*F* %	41 45.1	50 55.0	91	25 19.2	105 80.8	130
Failure of Recognition	*F* %	5 27.8	13 72.2	18	1 5.0	19 95.0	20

---

[13] There were a number of paradoxical instances of reported failure of recognition and correct matching. In part they might be accounted for by guessing. The number of instances was small.

ing should reduce the difference between them. Accordingly, we repeated the preceding experiment under incidental conditions.

Experiment 9: The Procedure of Matching—Incidental Learning. The experiments were done in groups of 4 to 8. As in Experiment 6, the stimuli were projected on a screen in a darkened room.

The task was that of judging the relative attractiveness of the forms and modes. In the nonunitary condition the *S*s stated whether the left (figure) or the right (mode) stimulus was the more attractive; in the U condition they indicated the relative attractiveness of the overall form and of its modes. An illustrative card was shown, as described in Experiment 8, in each condition. The *S*s were encouraged to judge spontaneously; the judgments were recorded on a rating sheet. A 4 sec. interval was introduced between exposures, during which the *S*s recorded their judgments. Special care was taken to insure that the *S*s would attend to the stimuli during the entire period of exposure, which was also 4 sec.

Recall was tested, after an interval of 3 min., by the method of matching. The procedure of testing was in every detail identical with that employed in Experiment 8.

The *S*s were 26 women students at Bryn Mawr College.

The results, which appear in Table 10, show that the incidental procedure markedly depressed the level of correct matching; indeed, matching of the non-U stimuli was not significantly different from chance. The differences between the incidental and intentional matchings, for the U and non-U stimuli, were significant, respectively, at $p = .03$.

At the same time the U stimuli were matched with a considerable degree of accuracy, one that exceeded significantly the level of non-U matchings. The mean number of correct U matches was 4.1, of correct non-U matches, 1.7 ($p = .01$). Indeed, the incidental matching of U stimuli was slightly superior to the intentional matching of non-U stimuli, a further indication of the effectiveness of the U condition. We consider it of interest that the U group continued to match with a substantial degree of accuracy under the incidental condition.

We again tested for the recognition of forms in the manner described for the preceding variation. Recognition of forms was somewhat, but not strongly, below the levels obtained with the intentional procedure (see Table 11). In this situation, we find that in the majority of cases recognition of form went with *failure* to match correctly. This finding is of particular interest with reference to the U stimuli (see pp. 36–39). A reference to Table 11a shows that the disparity with the synthetic measures of Experiment 6 was now quite extreme. This result cannot, as we saw, be traced more than very partially to lower recognition. We have to conclude that matching was to a considerable degree independent of recognition.

The comparison of these results with those of Experiment 8 provides particularly striking evidence of the difference between the U and non-U conditions. The levels of recognition were much the same in the two experiments, but they differed markedly in matching. Coherence of recall is not a function only of perceiving the requisite data; the perceptual relations between the data are decisive.

This variation establishes that the difference between unitary and nonunitary stimuli holds under conditions of incidental learning.

Experiment 10: Recognition of Pairing. The procedure called for the recognition of the original unit or pair. This mode of testing, which reduces to a minimum dependence on recall, should provide the clearest evidence concerning the difference between the stimulus conditions at the point of trace formation.

The experiment was done under group conditions, with 5 *S*s in each group session. The stimuli were projected on a screen in a darkened room, as in Experiments 8 and 9.

Familiarization with the forms and modes preceded the step of pairing. Here, we followed in all details the procedure described in Experiment 7. A test of recall of the single forms and modes followed the familiarization trials.

Thereupon, we proceeded to the joint presentation of forms and modes, under conditions identical with those prevailing in Experiment 5, except for the projection of the stimuli. The *S*s were assigned to the two main conditions, and were given intentional instructions appropriate to the U or non-U series. An illustrative stimulus during the instructions demonstrated the task in each condition. The test of recognition, which will now be described, came 3 min. following the exposure of the last stimulus.

a. Recognition test. For this test each form was reproduced on a separate page with 5 different modes. For the U condition, the form was drawn with each of the 5 modes on the contour; for the non-U condition, the continuous form was paired with each of the same 5 modes. One of the five pairings of form and mode on each page was correct. The two sets of 5 modes were selected so as to be of approximately equal difficulty, on the basis of the data of Experiment 6. (Each form was paired with 5, instead of with the entire set of 10, modes in order to reduce interference at the time of test.) There was a total of 10 pages, one for each stimulus.

The order in which the stimulus appeared in the recognition booklet varied from *S* to *S*, as follows. The recognition stimuli were divided into five sets of two each. They were rotated in pairs, so that, for some *S*s, each pair was the first in the recognition series. This procedure will permit us to determine the level of recognition of the entire set of stimuli when they appear early in the series, or before interference due to the method of testing has set in.

There remains to be mentioned one further detail of importance. The order of the stimuli was also shifted during learning, and in a fixed relation to the positions they had in the recognition series. The relation of the orders was as follows:

Recognition Order	1	2	3	4	5	6	7	8	9	10
Presentation Order	4	7	6	9	1	5	8	10	3	2

We call attention particularly to the fact that the first stimulus in the recognition series was always fourth in the learning series, and the second was always seventh in the learning series. The first pair of stimuli was thus by no means in positions most immune from interference during learning.

The recognition stimuli described above were lithographed and prepared as booklets consisting of 10 pages. The *S*s were instructed to check the correct pairing on each page; they were not allowed to return to an earlier page subsequently or to correct an earlier choice.

The *S*s were 50 women undergraduates at Douglass College, Rutgers University, of whom 25 were in the U condition and 25 in the non-U condition.

b. Results. Recall of the separate modes and forms following familiarization was identical in the U and non-U groups. The total mean recall score was 18.8 in each group. This value is closely similar to those obtained in Experiment 7; again, the recall of modes was slightly superior.

Recognition of pairing was substantially and significantly more accurate in the U group. *S*s in the U group correctly recognized 77% of the stimuli; the non-U group recognized 47% of the stimuli ($p = .00005$). The median of correct U recognitions was 9, of non-U recognitions 5. The range of correct U recognitions for the several stimuli was 14–24, of non-U stimuli 7–15. Each U stimulus was superior in recognition to the corresponding non-U stimulus. Figure 4, which summarizes the last result, plots the recognition scores of the stimuli in both conditions.

Further analysis showed that the position of the stimuli in the recognition test affected accuracy. Comparing the first and last two positions, we obtain mean scores in the U series of 88 and 72%, respectively; the corresponding non-U scores were 48 and 40%, respectively.

It is of particular interest to establish the level of recognition of the *first* stimulus in each recognition test. Since the stimuli were rotated in pairs, there were only five different stimuli in the first position. In the U series, these stimuli were recognized in 96% of the cases (by 24 out of 25 *S*s); the corresponding value for the non-U stimuli was 52%. It seems noteworthy that recognition of the first stimuli in the recognition series was nearly perfect in the U series, particularly when we consider that they oc-

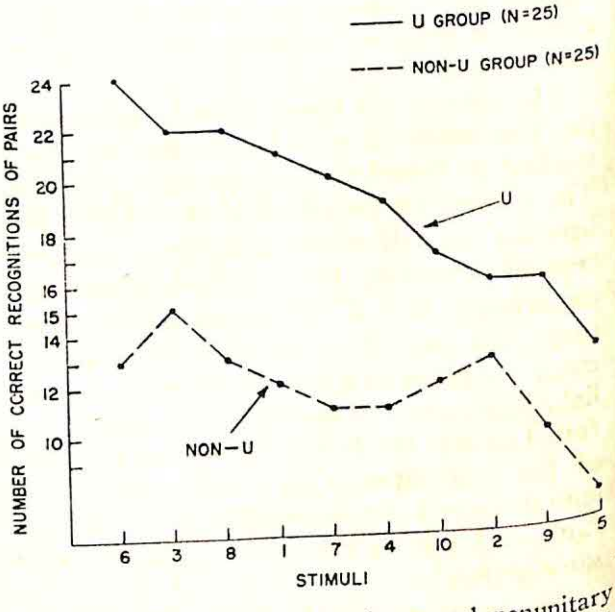


FIG. 4. Recognition of unitary and nonunitary pairs. This figure plots the number of correct recognitions of each stimulus in the unitary and nonunitary series. The stimuli were ordered in accordance with decreasing scores in the U-series. The stimuli are numbered as in Figure 3.

cupied the fourth position in the learning series, one that does not confer immunity from interference. We will consider the possible significance of this finding subsequently (see p. 38).

Figure 4 shows a definite correspondence between ease of recognizability of the individual stimuli in the two conditions. The rank order correlation between the U and non-U series was .70. Despite the systematic differences between the two conditions, the difficulty of pairing a given stimulus retained a substantial constancy.

These findings establish that the advantage of stimulus unity persisted when the contribution of recall proper was reduced to a minimum. We take this as evidence that stimulus unity favored the *formation* of functional connections.

THE FIGURE–GROUND RELATION

We have studied the role of two relations: of constitution and of the pairing of spatially separated heterogeneous units. The former we took to be an instance of unity; the latter, of duality. As mentioned earlier, there are most likely other sources of unity and duality. What follows is an attempt to explore the relevance to the present investigation of another relation—that of figure and ground.

The role of the figure–ground relation in the formation of associations has not been studied or raised to the level of a problem. The reason is a paradoxical one. There has been no realization that, considered as a perceptual situation, the classical association experiment has a highly particular structure. The stimuli of the association experiment are given as pairs; pairedness is established by means of contiguity. Analysis has failed to note the following regular property of the association experiment: the paired stimuli have been universally of coordinate rank, each having the status of *figure* on the same ground.

The pertinence of the figure–ground relation to the present inquiry is quite clear. Figure and ground constitute a unit, but they are segregated subparts, differing in phenomenal and functional characteristics. A study of the coherence between figure and ground should also throw light on a question arising from the preceding experiments. The unitary and nonunitary stimuli of the preceding section differed with respect to spatial *proximity*: form and mode coincided in locus in the one case, but were separated in the other case. Therefore, the question arises whether the results obtained were a function of spatial proximity alone. An investigation of the role of the figure and ground, which involves a change of proximity relations, should clarify this question.

Experiment 11: The Relation of Figure and Ground

The general technique of this investigation can be readily adapted to the study of coherence between figure and ground. The forms of Experiment 5 were drawn in thick continuous contour. The modal properties filled the entire background, including the inner area of the form. Figure 5 provides an illustration.

The instructions, as in the previous experiments, were to study each card so as to be able to reproduce the form and its background. The conditions of learning were identical with those of Experiment 5. A test of free recall was employed. There were 24 *S*s, 12 men and 12 women. The results appear in Table 12.

1. The mean level of recall, of both men and women, was somewhat but not significantly lower than in the U and non-U conditions of Experiment 5 (see Table 4). The decrease was greater for men than for women.

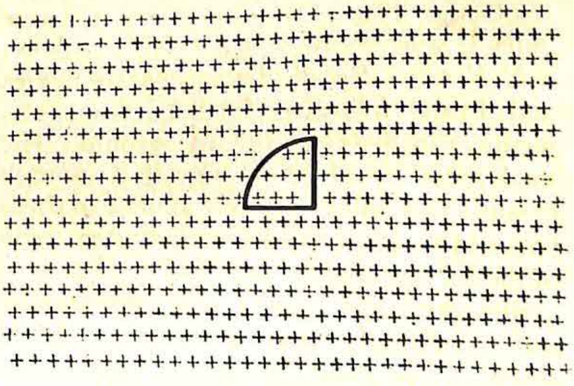


FIG. 5. A sample stimulus for the figure–ground condition.

TABLE 12

Recall Scores of Figure—Ground Stimuli
Experiment 11: Free Recall Test
($N = 24$)

	1 Correctly Paired Recall +(FM)	2 Incorrectly Paired Recall −(FM)	3 Recall of Form Alone	4 Recall of Mode Alone	5 Mean of Total Recall
M	3.3	2.3	1.8	1.5	8.8
%	37.1	25.7	20.5	16.7	

2. Correct pairings were significantly lower than in the U condition of Experiment 5 ($p = .006$), and insignificantly different from the non-U condition of Experiment 5 (see Table 4). The values of the index of pairing accuracy (see Table 6) place this condition unambiguously with all others that were nonunitary.

We conclude that the figure–ground organization acted as a *segregating* condition for purposes of association.

The results also have a bearing on the role of proximity, the investigation of which was another object of this variation. The proximity of form and modes approximated to that prevailing in the unitary condition, while differing markedly from the nonunitary condition. Nevertheless, the results were near to the latter, while departing strongly from the former condition. While this finding does not prejudge the general issue of the role of spatial proximity in association, it does suggest that the variations of proximity employed in the present investigation were without effect. The conclusion follows that the differences obtained between the unitary and nonunitary series are traceable to other sources, and that the relations of proximity were effective only in so far as they mediated other relations.

Further procedures suggest themselves that might clarify this first result concerning the role of figure–ground articulation. One obvious modification of the present procedure would be to exclude the modal stimuli from the inner area of the contour. It would be of particular interest to go one step further, and eliminate the continuous contour altogether, as illustrated in Figure 6. Part of this stimulus distribution is identical with the constitutive unitary stimulus, but the results should continue to diverge, in consistency with what we have found.

It would also be instructive, in further investigation, to fill the area enclosed by the contour with one set of modes, and the outer ground with a different set of modes.

The Relation of Inclusion

We now turn to another relation—to the inclusion of one visual form within the contour of the other. It is worth considering in its own right; in addition, it should throw further light on the role of spatial proximity under the present conditions. Finally, we will see that this variation stands in an interesting relation to Experiment 11.

Experiment 12: The Included Condition

The forms of Experiment 5 were drawn in continuous outline, and only the inner

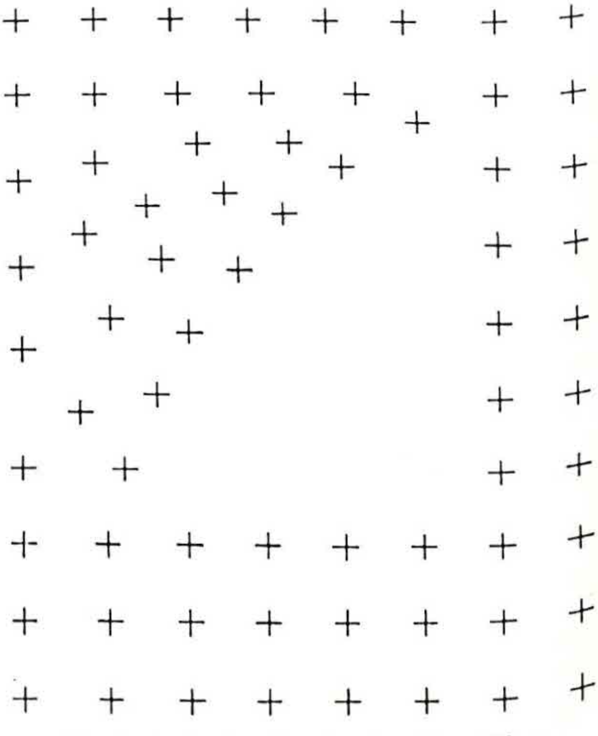


Fig. 6. A further figure–ground condition.

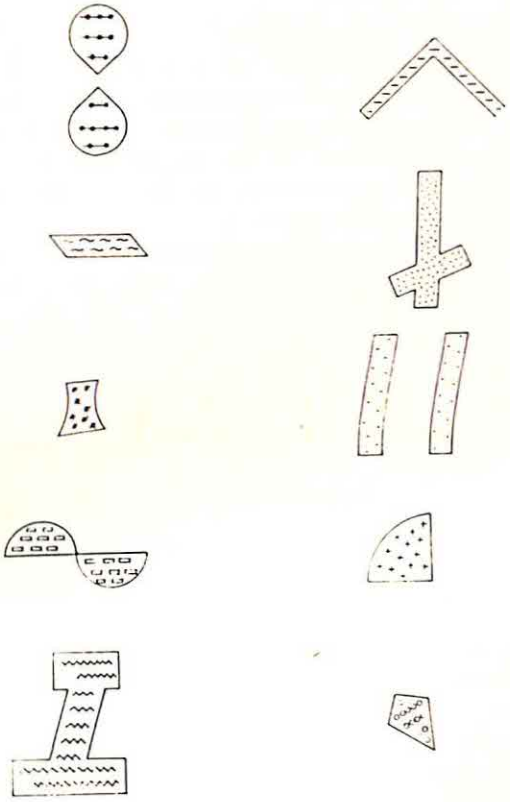


FIG. 7. Stimulus series for the included condition.

area of the contour was filled with the previously paired modes. Figure 7 illustrates the procedure. Thus, form and mode of Experiment 5 were retained, as well as the pairings.

The stimuli are reproduced in Figure 7. Each stimulus was again drawn on a white card (4 in. × 6 in.). When filling the area of a form with a modal property, care was taken to prevent the latter from touching the contour. Also, the modes were distributed, where possible, somewhat irregularly, to prevent them from outlining the form (and thus duplicating the earlier U condition). Finally, the modes filled the greater part of the area.

The *S*s were instructed to anticipate a test of free recall, during which they were to reproduce each form with its included modes. For purposes of explanation, an illustrative stimulus was prepared and shown during the instructions. In all other respects the procedure was identical with that of Experiment 5. The test employed was that of free recall. The *S*s were urged to reproduce each stimulus completely, to guess when not certain, and to reproduce the form or modes alone when they were unable to recall completely.

The *S*s were 12 men and 12 women.

Results. Total recall was somewhat, but not significantly, less than in Experiment 5, as Table 13 shows. Accurately paired recalls were 51.7% of all recalls, a value exactly between the U and non-U levels of Experiment 5 (see Table 4). It might therefore appear that the procedure of inclusion created an intermediate condition. In what follows it will be seen that this would be an incorrect conclusion.

The results contained a complicating circumstance. There were considerable differences among the stimuli in accuracy of joint recall. An examination of the stimuli raised the suspicion that they were not phenomenally of the same kind, although they had been prepared in accordance with a consistent and prescribed procedure. Some of the included modes appeared to "cohere" with their form far more than others. If this surmise should be correct, the stimuli did not constitute a homogeneous set.

To check on this possibility, we proceeded as follows. Six judges, two instructors in psychology and four graduate students, were shown individually the set of 10 stimuli spread out in a random (and constant) order. The judges, who had no knowledge of the experiment, were asked to select one stimulus the inner area of which "cohered" with the form, and another that lacked such coherence. The instructions did not attempt to define the basis of the division beyond the statement that form and mode "went better" together in some instances than in others.

TABLE 13

RECALL SCORES OF INCLUDED STIMULI
Experiment 12: Free Recall Test
($N = 24$)

	1 Correctly Paired Recall +(FM)	2 Incorrectly Paired Recall −(FM)	3 Recall of Form Alone	4 Recall of Mode Alone	5 Mean of Total Recall
M	4.4	1.6	1.6	1.0	8.5
%	51.7	18.5	18.5	11.2	

Thereupon, the judges were requested to divide the entire set into two equal halves on this basis. Finally, the judges ordered the stimuli in each half according to degree of coherence and lack of coherence, respectively. We thus obtained from each judge a ranking of the entire set of stimuli on a dimension of "coherence."

We scored the stimulus selected by each judge as most coherent 10, the next 9, and so on. The scores assigned to each stimulus by all six judges were then summed. Thus, we obtained an ordering of the stimuli in accordance with a phenomenal criterion of coherence.

There was a very substantial agreement among the judges. The mean rankings of the stimuli, in the order in which they are reproduced in Figure 7, were: 2.0, 7.7, 3.7, 3.0, 6.2, 8.7, 7.5, 7.7, 2.7, and 6.0.

We then divided the stimuli into two halves of five each, in accordance with their mean rankings, and calculated the respective recall scores. Let us call the upper half "coherent" (or C), and the lower half "not coherent" (or −C). (The C stimuli were: 2, 5, 6, 7, and 8; see Figure 5.) For purposes of comparison we grouped similarly the scores of the corresponding stimuli in Experiment 5, both in the U and non-U conditions. The results appear in Table 14.

There was an impressive correspondence between the recall scores and the—independently obtained—judgments. First, the total recall of stimuli judged coherent was significantly superior to the recall of the less coherent stimuli (see Table 14, row 5). Further, this superiority was almost completely due to greater accuracy of joint recall. Correctly paired recalls comprised 67.2% of all recalls among the C stimuli, and 27.5% among the others. By the Wilcoxon test, the difference in absolute number of pairings is significant at far beyond .01. The difference is also significant when the proportion of correct pairings to total recall in every category is the score; by the sign test this difference is significant at beyond .002.

For comparison we have grouped the results of Experiment 5 similarly. Since the basis of division is irrelevant to this variation, it is not surprising that there were no appreciable differences between C and −C stimuli. Indeed, accuracy of pairing of −C stimuli tended to be somewhat higher in Experiment 5.

The principal finding is the close resem-

TABLE 14

COMPARISON OF COHERENT AND NONCOHERENT STIMULI

	Experiment 12		Experiment 5			
	Included Stimuli ($N = 24$)		Unitary Stimuli ($N = 24$)		Nonunitary Stimuli ($N = 24$)	
	C	−C	C	−C	C	−C
1. Mean Correct Pairings: +(FM)	3.5	0.9	2.9	3.4	1.9	1.8
2. Mean Incorrect Pairings: −(FM)	0.7	0.9	0.8	0.5	1.7	1.5
3. Mean Recall of F Alone	0.5	1.0	1.0	0.7	0.5	0.5
4. Mean Recall of M Alone	0.5	0.5	0.1	0.3	1.1	0.7
5. Mean Sub-total Recall	5.2	3.3	4.8	4.9	5.2	4.5
6. Mean Total Recall	8.5		9.7		9.6	
7. Percentage of Correct Pairings (1 ÷ 5)	67.2	27.5	60.9	70.1	37.1	39.3
8. Percentage of Incorrect Pairings (2 ÷ 5)	13.6	26.2	17.4	10.3	33.1	34.6

blance, with respect to accuracy of pairing, between the C stimuli of the present experiment and the corresponding U stimuli of Experiment 5, and the resemblance between the −C stimuli of the present experiment and the corresponding non-U stimuli of Experiment 5.[14] The values of the index of pairing accuracy (see Table 5) offer further convincing support for this statement.

Includedness per se is thus not a condition favoring joint recall. This conclusion receives support from an earlier unpublished experiment by Celeste McCollough and the writer. The association between pairs of heterogeneous nonsense figures was compared when they were spatially separated and when one appeared inside the contour of the other. No differences of association were found between these conditions.

Again, the results bear on the role of proximity. With proximity (of form and mode) equalized, there were nevertheless marked and consistent differences between the two subgroups that were discriminated on a phenomenal basis. Proximity clearly cannot account for this result. Further, we have also found that, with respect to pairing, the C stimuli of this series were similar to the U stimuli of Experiment 5, and that the −C stimuli resembled the non-U stimuli. This result suggests that the differences in pairing between the U and non-U stimuli were also not a function of proximity. (For a further discussion of this point see pp. 31, 34.)

We also note that, as a consequence of the relation of inclusion, all stimuli of this experiment were phenomenally of the "one" variety. If so, membership in a "single" perceptual constellation is not a condition that favors pairing. We need to distinguish between the property of "singleness" and the quality of unity. The difference between the unitary and nonunitary stimuli of the earlier experiments is therefore not adequately described in terms of the difference between "one" and "two" either. Indeed, it can be readily shown that neither proximity nor numerical unity are decisive for figural unity.

What, then, is the perceptual difference between the stimuli we have called "coherent" and "not coherent"? To us it seemed that the form and contour of the stimuli judged coherent participated in the formation of a unitary *surface.* In these cases the filled area was seen as belonging to the form. In the other cases the inner area was seen simply as filled with forms of a particular kind; the contour and the enclosed shapes had distinct identities. If this interpretation is in the right direction, the C stimuli were phenomenally and functionally unitary, in contrast to the −C stimuli, and the present results provide independent evidence concerning the effect of perceptual unity on pairing.

It is a question for separate investigation to establish the stimulus conditions of surface formation in this case. The present study provides only tentative hints. It seemed to us that the less individual the modes were as forms in their own right, the better they lent themselves to this effect; thus, dots and stippled areas were judged predominantly coherent. It is also probable that the size of the included forms and the uniformity of their distribution, as well as their figural relation to the including form, play a part.

It also follows that the specific sources of phenomenal unity are diverse. According to the interpretation here offered, in the case of the C stimuli, includedness mediated a particular organization, namely, the formation of a surface, just as the relation of constitution mediated the phenomenal unity of the stimuli of the preceding experiments. It would appear that the complete spatial coincidence of two properties is one condition of unit formation. Correspondingly, nonunity, under certain conditions, may be achieved by spatial separation, and also by failure of surface formation.

The present experiment stands in an interesting relation to Experiment 11. Viewed in a certain way, they are identical in this respect: both have continuous forms with modes filling the inner area. The difference between them is equally clear: in the present experiment the modes were confined entirely within the contour, whereas in Experiment 11 they extended into the background. The following comparison between them seemed, therefore, worth attention.

---

[14] As mentioned above, the C and −C stimuli differed also in general level of recall. We have no explanation for this result. The combined condition may have accentuated the difference between the two subsets. We found some evidence of the same trend in Experiment 4.

We have just identified two subsets of stimuli—coherent and not coherent—that differed markedly. Let us now observe the performances on the corresponding subsets of the figure–ground condition. The results appear in Table 15.

The difference between coherent and not coherent stimuli disappears completely when the modal components extend into the ground. The finding that the interaction between a contour and included modes is abolished when the latter are also part of the ground makes good sense from the standpoint of perception. When the modes are part of the ground, those falling within the contour lose their specific relation to it, and become a mere continuation of the ground. Thus, we have established a difference between the relation of inclusion and the figure–ground relation.

This investigation has provided further confirmation of the role of stimulus unity. The difference between the more and less unitary stimuli corresponded to those of the earlier experiments. The specific grounds of unity were different in the present setting, but the consequences for joint recall were quite similar.

Experiment 13: A Variation of the Included Condition

As a check on the preceding results, Experiment 12 was repeated with a new pairing of forms and modes. The forms and modes were those of the preceding variation, but the pairings were now those of Experiment 1.

The procedure described above was followed without change. Again, the stimuli were judged by a (new) group of 6 judges for "coherence." The *S*s were 24 Swarthmore College students, 12 men and 12 women.

The results were in all essentials similar to those of Experiment 12. The new sets of C and −C stimuli differed markedly in total recall and in accuracy of joint recall. By the Wilcoxon test the difference in absolute number of pairings was significant beyond .01. Correct pairings accounted for 70.8% of all C recalls, for 46.5% of all −C recalls. The ratio of correct pairings to the total recall score of each *S* was also computed for each *S* in each category; by the sign test, $p = .03$.

TABLE 15

COMPARISON OF COHERENT AND NONCOHERENT STIMULI

	Experiment 11		Experiment 12	
	Figure-Ground Stimuli ($N = 24$)		Included Stimuli ($N = 24$)	
	C	−C	C	−C
1. Mean Correct Pairings: +(FM)	1.4	1.8	3.5	0.9
2. Mean Incorrect Pairings: −(FM)	1.0	1.3	0.7	0.9
3. Mean Recall of F alone	1.0	0.8	0.5	1.0
4. Mean Recall of M alone	0.8	0.7	0.5	0.5
5. Mean Sub-total Recall	4.2	4.6	5.2	3.3
6. Mean Total Recall	8.8		8.5	
7. Percentage of Correct Pairings (1 ÷ 5)	34.0	40.0	67.2	27.5
8. Percentage of Incorrect Pairings (2 ÷ 5)	24.0	27.3	13.6	26.2

These results actually underestimate the difference between C and −C stimuli. While the judges were consistent in differentiating between the extreme stimuli of this series, they were far more variable when it came to the 4 middle stimuli. The mean rankings of the latter were very close to each other, attenuating the effect. When we compared the three stimuli at each extreme, correct C pairings accounted for 70.6% of all recalls, and correct −C pairings accounted for 37.3% of all recalls. The results substantially confirm the findings of Experiment 12.

We have also asked whether the same modes and forms were judged coherent in Experiments 12 and 13. As might be expected, the modes proved more decisive. Three modes (dots, dashes, and stippled areas) produced high judgments of coherence when they filled the areas of different forms. As mentioned earlier, the lack of individuality of the modes, and the homogeneity of their distribution, was probably responsible for the impression of areal homogeneity, and consequently for surface formation.

Experiment 14: A Further Variation

The principal result of Experiments 12 and 13, namely, the difference that emerged between stimuli judged "coherent" and "not coherent," is susceptible of another interpretation. We have seen that the former generally contained modes of relatively homogeneous distribution. This suggests the possibility that modes belong to "coherent" stimuli duplicated the shape of the contour. If so, the advantage they had may not have required the participation of the boundary line. The modes in question would then represent a modified version of the earlier U stimuli, while the −C stimuli would correspond to the earlier non-U condition. Consequently, the interpretation advanced above would be incorrect. The present experiment was designed as a check on this possibility.

The stimuli of Experiment 12 were modified as follows. Each card consisted of a pair of separated stimuli. The stimuli at the left were the continuous forms of Experiment 12, those at the right were the arrays of modes which belonged to that form. The latter were grouped precisely in the way they were previously. Thus, the stimulus contents were identical with those of Experiment 12, except that the modes were now separated from the contour.

The *S*s were given free recall instruction. In addition, the instructions specifically called attention to the similarity of shape between the terms, as the following statement shows: "Each card will contain two figures. One will be that of a simple outline form at the left; and the other will be that of a number of smaller forms grouped together at the right. It will help you in remembering the pairs later to notice that the smaller drawings at the right have an overall shape quite similar to the one at the left. I want you to look at the drawings carefully so that later you will be able to draw from memory as many of the pairs as possible." A free recall test followed.

If similarity of form between contour and modes was responsible for the difference between the "coherent" and "not coherent" stimuli of Experiment 12, this difference should also appear under the present conditions. If the difference between the two sets of stimuli should disappear, the conclusion would be in order that "coherence" in Experiment 12 was mediated by the surrounding contour.

The *S*s were 7 men and 5 women students at Swarthmore College. The results, which appear in Table 16, contain the data for the stimuli which had been categorized in Experiment 12 as "coherent" and "not coherent."

Mean recall was closely similar to that obtained in Experiment 12. The proportion of correct pairings to recall, for the entire series, was somewhat lower; there was also a higher level of isolated recalls.

The C stimuli did show a trend toward more accurate pairing than the −C stimuli, but the difference was small and not significant; the effect was far weaker than in Experiment 12. There was no difference between C and −C stimuli in inaccurate pairing, in contrast to the results of Experiment 12.

The results do not convincingly support the view that the advantage of the "coherent" modes is to be traced to their form quality, and that this accounts for the findings of Experiments 12 and 13. It should be added that the present variation may have been weighted too one-sidedly against the surface formation alternative. The procedure of separating the continuous contour from the mode aggregates, it seemed to us, enhanced the perceived form similarity between them beyond what was the case in Experiments 12 and 13, in addition to which the instructions explicitly stressed this relation of similarity. Nevertheless, given our lack of knowledge concerning the interaction between a contour and its included modes, one must reserve a final conclusion.

We can, however, draw the lesser conclusion that the factor of proximity was not responsible for the differences obtained between the unitary and nonunitary stimuli of the earlier experiments.

INTERPRETATION AND DISCUSSION

We have studied an aspect of coherence that has not hitherto claimed attention. Theory and investigation in this region have placed in the forefront the role of conditions such as temporal contiguity and frequency. This investigation was concerned with a different question, namely, with the role of different kinds of stimulus relations in establishing coherence.

Accordingly, we compared the formation of functional connections between stimuli whose relations were varied in a number of ways. The stimuli were heterogeneous visual forms, and the relations between them were as follows: they were spatially segregated units; they belonged to the same visual contour, being in a constituent–whole relation; the boundaries of one surrounded the other; one had the function of figure, the other of ground. These conditions differed in important ways. More particularly, they could be grouped on the basis of the memory data in two categories. One set of conditions produced markedly superior connections, the other was more resistant to their formation. Table 17 summarizes the findings for the entire range of conditions.

The two major groupings include relations of considerable diversity. We proposed that the basis of division was the property of unity, that unitary stimuli were superior in establishing functional coherence.

What grounds are there for identifying a given stimulus distribution as unitary? To avoid circularity the decision, of course, must be independent of the memory data. We relied in the first instance on phenomenal evidence. The division of the stimuli on a phenomenal basis was straightforward. In some of the instances studied, special inquiry was unnecessary. Most, if not all, observers would agree to call a set of repetitive modes delineating a shape unitary, and the spatial separation of these from a continuously outlined shape dual (as in Experiments 1–10); at least they would agree that the former are more unitary than the latter. There would be similar agreement that the figure–ground distributions in Experiment 11 are less unitary than the forms delineated by modes. In other instances that were less clear, phenomenal judgments were systematically obtained for the purpose; this

TABLE 16

COMPARISON OF COHERENT AND NONCOHERENT STIMULI
Experiment 14: Free Recall Test

	C	−C
1. Mean Correct Pairings: +(FM)	2.2	1.3
2. Mean Incorrect Pairings: −(FM)	.5	.3
3. Mean Recall of F alone	1.0	1.4
4. Mean Recall of M alone	.7	.9
5. Mean Sub-total Recall	4.4	4.0
6. Mean Total Recall	8.4	
7. Percentage of Correct Pairings (1 ÷ 5)	49.1	33.3
8. Percentage of Incorrect Pairings (2 ÷ 5)	11.3	8.3

TABLE 17

Summary Table of Recall Scores of Unitary and Nonunitary Stimuli
(in %)

Unitary Stimuli			Nonunitary Stimuli		
Experiment	Correctly Paired Recall	Incorrectly Paired Recall	Experiment	Correctly Paired Recall	Incorrectly Paired Recall
1	76.2	7.4	1	49.2	31.4
2	77.8	22.2	2	40.7	59.3
3	74.6	25.4	3	48.5	51.5
4	88.7	0.9	4	29.3	37.8
5	65.5	13.8	5	38.1	33.8
12:C	67.2	13.6	12:−C	27.5	26.2
			11	37.1	25.7

Note.–The experiments included in this table employed the method of free recall, except for Experiments 2 and 3, which employed paired-associate recall.

was the case with the inclusion of modes in the area bounded by a contour (see Experiments 12 and 13).

Since the objective stimulus properties in most cases can be specified exactly, it is possible to relate them to phenomenal unity. Accordingly, we can say (*a*) that the relation of constitution is a condition of phenomenal unity, (*b*) that the spatial separation of nonsimilar forms is a condition of duality, as is also (*c*) the presence of hetergeneous forms in the relation of figure and ground. The stimulus conditions adequate to surface formation, as we have seen, await further investigation.

Implicit in this discussion is the assumption that unity (and duality) has its source in different specific conditions, some of which we have doubtless not mentioned. The question arises whether the general category of unity, when applied to instances as different as those here considered, and to others that were not considered, has a concrete content. It is not possible to specify at present an objective stimulus property shared by all instances experienced as unitary or dual, respectively. Would it therefore not be simpler to describe the results in terms of the objective characteristics of the stimulus and to forego reference to a property that can only be identified phenomenally? Such a procedure would leave entirely unexplained the fact that each set of diverse stimulus relations produced consistently similar results. Although the grounds of phenomenal unity require further investigation, they should not for that reason be ignored. It may well be that the phenomenal evidence is in advance of what we are at present able to formulate in other terms.

A further question will have to be faced in connection with the property of unity. There may be grounds for holding that any two stimuli, however heterogeneous, constitute a unit, when they are the only items in the field. Another way of stating the same point would be to consider that any pair of units may form a single pattern. It would follow that the stimulus conditions we are contrasting are all unitary, and that they differ mainly in degree of unity. If so, the question we have put would concern differences between more or less unitary conditions, or between units differing with respect to the autonomy of their subparts. It is of considerable importance to establish whether the difference in question is only one of degree, and we will consider this point subsequently (see pp. 39–43). However this issue is decided, the main outcome of this investigation remains unchanged. It has demonstrated that the coherence of data in memory is a function of their unit relation.

The findings are restricted to a particular modality—the visual—and to a limited set of properties within the modality, and therefore need to be verified in other settings. But they provide the basis for a general conclusion: whenever we can specify the relative unity of two stimulus distributions, we can predict their relative coherence.

In what follows we will explore further the conclusion just stated, including its bearing on the nature of associations.

Alternative Interpretations

The procedure we have followed was to relate differences of recall to given stimulus conditions. It is necessary, therefore, to establish whether stimulus properties other than those we have identified might have been responsible for the findings. This question has concerned us earlier; we will now consider it more systematically.

1. One obvious requirement of this study is that the respective stimulus conditions be equally perceptible or discriminable. In the light of the results already reported, this requirement can be modified as follows: differences in perceptibility must not favor the conditions that are superior in coherence. Experiment 6 demonstrates, we believe, that this requirement was met. There, we employed recognition (of forms and modes, respectively) as an index of perceptibility. (*a*) The main finding was that the nonunitary stimuli (in particular the modes) were somewhat more recognizable. This circumstance should favor the joint recall of nonunitary stimuli. Nevertheless, the unitary condition was superior with respect to joint recall. (*b*) It might be maintained that the issue concerns the perceptibility of the entire stimulus, or the perceptibility of the given relation. For example, was the constituent–whole relation more easily discriminable than the relation of adjacency between heterogeneous forms? Such threshold differences could determine the memory results.[15] In this connection we would cite the "synthetic pairings" of Experiment 6, or the instances of *joint* recognition of the form and mode of a given unit or pair (see p. 14). Such joint recognition was more frequent in the nonunitary series. We take this result to mean that the threshold of perceptibility was lower in the nonunitary series. (*c*) Experiment 7 provides evidence in the same direction. The procedure of familiarization there employed raised the recognizability of the forms and modes to a maximum and also equalized it for the two conditions. The persistence (and increase in magnitude) of the experimental effect under these circumstances is convincing evidence that differences of discriminability were not responsible for the results.

2. The rationale of this study requires that the conditions be equivalent in content, or in "amount" of information. No finding is more clearly established than that it is more difficult to master more material than less. When two otherwise similar tasks differ in ease of mastery, the question immediately arises whether they differ in "amount to be learned."

a. The assumption of such a difference may take the form of asserting that the unitary series consisted of single stimuli, and the nonunitary series, of pairs; or that the latter contained twice as many stimuli as the former. This formulation begs the question. It ignores the fact that each condition contained forms and modes, that these were heterogeneous, and that one could not be deduced from the other in either condition. The assertion is actually an outcome of the correct intuitive inference that the unitary condition is the easier, and of the unexamined assumption that the responsible difference is one of amount of material.

b. The following is a more serious version of the same point. The nonunitary stimuli consisted of pairs, each with its form and mode. The *a* term was a particular form delineated by a continuous mode; the *b* term also had a particular shape, that of a straight line, and was composed of modes. Thus, each pair contained two forms and two modes. The unitary stimuli, however, eliminated the continuous mode and the linear shape.

---

[15] The writer is grateful to E. J. Gibson for calling attention to this point.

Two points might be made in answer. First, the nonunitary *S*s knew from the start that the stimuli would consist of continuous forms paired with lines; they did not need to learn about this feature of the task. It must be admitted, though, that this is not a conclusive reply; the fact remains that the nonunitary series contained data absent from the unitary series. Second, and more weighty, is the consistent finding that the conditions were very similar in content recall. This result goes counter to the interpretation that a quantitative difference of material was responsible for the observed effect.

3. A more plausible objection is that the unitary stimuli were less homogeneous, and therefore less vulnerable to mutual interference. Each condition had a *series* character. In particular, the nonunitary set contained two well-defined series—continuous forms and straight lines. The members of each of these series, by virtue of similarity, may have been more exposed to interference than those of the unitary series. Although this homogeneity did not impair recall of content, as we saw earlier, it is quite possible that it contributed to associative interference.

It was in this connection that we performed a control experiment employing the procedure of familiarization (Experiment 7). This experiment provided strong evidence that differences of intraserial interference do not account for the results. Familiarization raised the recall of forms and modes to a near maximum prior to the main task, equalizing the availability of the items in the respective conditions. This could only have the effect of reducing the importance of any differential interference at the time of learning. If interference had been responsible for the results, differences between the two series should have been reduced. The opposite was the case. (It should be added that if we had obtained evidence for the differential role of interference, this would not have necessarily resolved the issue. It would still remain to be decided whether the interference in question was a cause of the experimental effect, or a consequence of the unit relation.)

Here also should be mentioned the experiments involving the relation of inclusion. The presence in the series of dense fields of modes, in accordance with the argument, should have made for very considerable interference; the clustered modes were as homogeneous as their counterparts in the nonunitary series of the earlier experiments. The results support the interpretation that it was the relation of form to modes which exerted the decisive effect.

Finally, it seems appropriate to caution against too ready an acceptance of the relevance of an interference interpretation for data such as these. It seems to us more tenable that the poorer performance in the nonunitary condition was a direct consequence of the greater resistance these stimuli offered to the formation of associations, and possibly of their lower resistance to disintegration.

4. Unitary and nonunitary stimuli differed with respect to the spatial proximity of forms and modes. If we had to rely on the data of Experiments 1 to 5 alone, it would be a tenable conclusion that spatial proximity was responsible for the results obtained. Accordingly, we altered the proximity of forms and modes in Experiments 12 and 13. Although these variations equalized the proximity relation for all members of the series, they produced two sets of stimuli differing appreciably in pairing. One of the sets gave pairings quite comparable to those obtained with the nonunitary stimuli, despite the evident differences between them in proximity. Finally, the results of the figure–ground variation of Experiment 11 resembled closely those found with the nonunitary stimuli, although the two again differed in proximity.

We thus found that conditions identical in spatial proximity differed in joint recall and that others differing in proximity were equivalent in this respect. Proximity cannot account for these results. Whatever the effect of proximity per se might be, other circumstances must be in large part responsible for the present findings. The interpretation we propose is that spatial proximity was of consequence in this investigation only insofar as it affected the property of unity, and that differences of proximity were otherwise without appreciable effect.

5. A related question is whether the contrasted conditions differed with respect to the temporal character of the stimulation. In the unitary case form and modes were most likely perceived simultaneously or nearly simultaneously, but they may have been apprehended in a more successive way when they were spatially separated. Indeed, one finding reported earlier may seem to point in this direction (see Experiment 3a), and to suggest that a division of attention played a part when the stimuli did not occupy the same position.

Differences of temporal stimulation, we believe, cannot account for the results. First, in the included and figure–ground conditions, form and modes occupied the same regions, a circumstance that should have reduced successiveness, but the results do not go in the presumed direction. Second, there is no convincing reason for the belief that present differences of temporal successiveness could be responsible for the order of effects obtained in experimentation of this kind. Finally, successiveness of apprehension, to the degree it played a part, may itself have been a consequence of the perceived lack of unity.

6. It is reasonable to ask whether unitary stimuli lent themselves more easily to mnemonic devices. A unitary stimulus might be more likely to resemble familiar objects. Also, it might be verbally describable in a more concise way than a pair of stimuli; thus, the designation "dotted circle" is perhaps easier to recall than "circle and dots." If so, the experimental effect would be only indirectly an effect of the perceptual properties in question, and a direct consequence of the intervening verbal operations.[16]

Our observations showed that labelling did occur with some frequency, though by no means uniformly. We have no evidence that the conditions differed systematically in the manner suggested above; it is unlikely that this was the case. (*a*) Perhaps the most convincing evidence against this interpretation comes from Experiment 9, which employed a procedure of incidental learning. The latter should have decreased the use of labelling in both conditions, and consequently the magnitude of the experimental effect. This was not what we found. (*b*) This interpretation cannot readily account for the differences between the "coherent" and "not coherent" subgroups of Experiments 11 and 12.

7. Finally, it may appear that the findings can be interpreted in terms of Thorndike's factor of "belonging." It is clear to us that this is not the case. The nonunitary conditions of this study conformed completely to the requirement of belonging, in the sense of Thorndike, fully as much as the unitary conditions. Thorndike (1932) restricted belonging to the realization that "this goes with that," expressly asserting that it referred to "nothing logical, or essential, or unifying. . . ." (p. 72). It would therefore be confusing to identify belonging with the relation of unity.

Let us summarize the central issue of this section. We have found that a unitary stimulus joins its parts more intimately, or that one can go, in recall, from one aspect of it to another more easily than one can from one unit to another. What, we may ask, is the selective principle responsible for this advantage? Is unity per se a condition of coherence? Is it an irreducible datum, or can its effects be derived from other determinants? More particularly, can these effects be referred to conditions that are only concomitant with unity, and which themselves lack organizational character? Had we demonstrated that these other conditions could account for the results, the significance of the investigation would have shifted appreciably. To be sure, it would still be a matter of interest to have shown that unity implies the presence of other characteristics important for learning and recall. But in that case the relation of consequence would be between the data of recall and properties such as intraserial homogeneity or perceptibility; the condition of unity would have become merely an index of the presence of these other properties.

---

[16] One may ask, though, whether the assumption that the one verbal phrase is more effective than the other does not presuppose the very principle that is in question.

It is hardly our purpose to suggest that the consequences of unity are not susceptible of further analysis. The preceding examination demonstrates that certain particular interpretations are not valid, and raises the presumption that the search for an answer will have to take relational properties of the stimulus into direct account.

Sources of Perceptual Unity

We would like to understand better the grounds for the fact that a unitary stimulus joins its parts more intimately, thus producing characteristic effects in memory. An answer requires a more detailed study of the stimulus. The following is an attempt to re-examine the stimulus relation that was here most extensively studied, that of form and mode on a contour line.

It should first be understood that the relation of form and mode is not an instance of any of the grouping principles described by Wertheimer. These principles referred entirely to what we have here called the property of form. To be sure, Wertheimer often illustrated these by means of patterns of dots, the latter functioning as modes. His concern, however, was solely with the particular grouping that was realized out of the alternatives that were theoretically possible. For this purpose the character of the modes was of hardly any account. In contrast, the question we have raised concerns precisely the relation between the form that is generated by modes that "carry" the form.

It is also important to observe that the relation of form to mode is not a usual instance of a part–whole relation. A mode is not part of an encompassing form in the same sense that a side is part of a square. This follows from the fact that the modes contribute to the form primarily by virtue of their position; they have the function of points in generating a form. The specific shape of the mode is far less relevant for this purpose. This can be easily demonstrated. Different modes can serve to delineate the same form (see Figure 1); in these cases there is no difficulty in identifying the constant form. Conversely, the same modes can delineate different forms; and again one notes readily the identity of modes and difference of forms.

We have called the relation of form and mode one of constitution, since the modes are strictly constituents of the form. The relation is also one of *dependence;* one aspect of the stimulus requires the other. It is probable that stimuli identified as unitary are as a rule dependently related, either in the sense that one cannot occur without the other, or because they alter each other. In that case the distinction between a unit and a duum would be clearly defined: the latter would involve nondependence and absence of (or extremely weak) mutual modification.

At this point we may venture a further step, based on the observation that form and modes were identical in locus. The form–mode relation may be a special case of a general principle which can be stated as follows: *The spatial coincidence of any two properties or aspects is a sufficient condition of unity.* Actually it is not certain that the form–mode relation is one of coincidence. The following would provide a more decisive test of the principle stated above. Let one stimulus series be a set of colored forms, each on a white ground. That is to say, the area of each form is of a homogeneous and distinctive hue. Let the memory task be that of recalling (or matching) the hue of each form. One can construct a parallel series of identical forms, with the ground of each in homogeneous and distinctive hue, while the inner areas of the forms are white. The first series is an instance of coincidence of color and form, in contrast to the latter. In terms of the principle stated above, the first stimulus series should be superior in matching form with hue.[17]

---

[17] Experiments employing this procedure will be reported subsequently.

It would be an error to consider the principle of spatial coincidence an instance of a law of proximity. The latter (for which we have found no support under the conditions of this investigation) presumes a continuous relation between distance and a given psychological function. We have seen, though, that there develops a strong difference between complete spatial coincidence and a condition that departs only slightly from it (compare Experiment 1 with Experiments 11 and 12).

We have not attempted to explain why the relation in question favored functional coherence. Given the restricted degree to which form and mode alter each other, the strength of the obtained effect is indeed impressive. The following remarks are intended only as suggestive.

First, we note that the modes of this investigation had two properties: they were forms, and they functioned as points in generating a form. To us it seems deserving of investigation to establish whether a *double function* is a condition favoring coherence. At present, there is no evidence on this point.

Second, it may be going too far to say that form and mode do not interact. Some modes may well be more compatible with a given form than other modes (see Figure 8). Thus, a circular mode (Figure 8a) tends to blur an angular form such as the triangle, especially at the apex points, but is less affected by angular modes (see Figure 8b). At the other extreme, there is an outstanding relation of fittingness between a form and a mode of the identical form (see Figure 8c). At the same time, modes with least individuality as forms in their own right, such as dots and continuous lines, least disturb the form property (see Figures 8d and 8e), and are suitable for the delineation of all shapes.

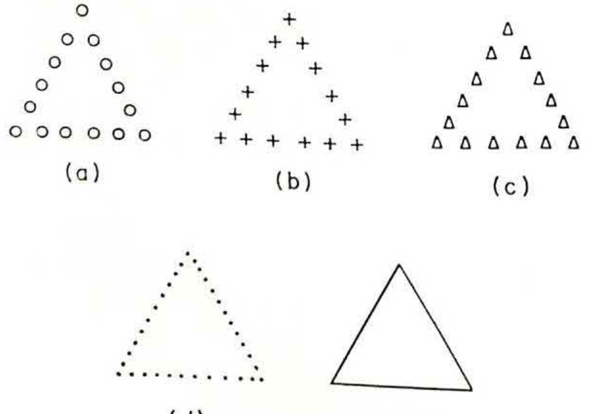


FIG. 8. Compatibility of form and mode.

Although there is no advance relation of requiredness between form and mode, they may be more or less compatible once they are paired. Thus, compatibility may introduce a more intrinsic relation between form and mode. Also, the possibility arises that in the unitary condition the *S* is more directly centered on the relation between form and mode. For example, he may be sensitive to the fact that a curved form is carried by angular modes. This formulation does not explain away the obtained difference between the unitary and nonunitary stimuli, since a difference of sensitivity of the kind here assumed would itself be a function of the property of unity. In addition, it is doubtful whether such distinctions can adequately account for the order of the obtained effects.

Finally, it is a question for separate investigation whether the more usual category of part–whole relation produces effects similar to those under discussion. The interpretation put forward in this study, namely, that perceptual unity is a condition of coherence in recall, requires an affirmative answer.[18]

This discussion may serve to illustrate the fruitfulness for the present problem of an analysis of the stimulus. (*a*) The constitutive relation was distinguished from the part–whole relation and from the conditions of perceptual grouping hitherto studied. (*b*) We have shown that the constitutive relation is one of dependence, and suggested that dependence may be a property of unitary stimuli generally. (*c*) A principle of spatial coincidence was stated as one condition of phenomenal unity. (*d*) The question was raised whether the presence of a double function favors coherence. (*e*) The relevance of the part–whole relation to the present problem was described, and shown to furnish a test of the general interpretation growing out of this investigation.

---

[18] Experiments dealing with this question will be reported subsequently.

Perception and Association

This investigation deals with the relation between perceptual properties and coherence as evident in the data of memory. It therefore makes contact with the older problem of the relation of perception to association. The major theories of association have at this point followed different, indeed opposed, assumptions. These we will now examine, in order to see their relation to the present problem.

1. What one may call associationistic doctrines of association hold to a sharp separation between perception and association. This is a direct consequence of the historical fact that these doctrines started with the assumption that psychological contents were sense data, which had the status of independent, irreducible elements. This starting point dictated the scope and conception of the associative process. In the first place, it was to association that one had to turn for the explanation of order in mental life. Sense data, being discrete, lacked order; they had to be combined into stable sequences, and this was the function of association. There were thus two major problems for this psychology: the identification and description of the sense data or elements (and eventually the study of the physical energies that are their adequate stimuli), and the investigation of the manner in which they become joined. The fundamental task of associationistic inquiry was to account for the joining of one sense datum to another, the process that was held responsible for the constitution of psychological objects, and in general for the principles of order that unorganized sense data lacked.

Given the sense data thesis, a number of consequences followed for the interpretation of association as a process. If the function of association was to join data that are initially unconnected, then the noting of sense data was one kind of event, the association of them another. On this basis an association was conceptualized as the formation of a bond or pathway between unchanging contents, and independent of the character of the contents. Some associationistic theorists, such as John Stuart Mill, were not content with this formulation, suspecting that association often produced a more intimate joining of contents. It was in this connection that the notions of "fusion" or "mental chemistry" were brought forward, but the problem was not to be faced until the appearance of the gestalt theory of perception.

This sharp separation between the data and their association served as the theoretical basis for the experimental study of associations. Let it be noted that, in terms of the present discussion, an association has been historically conceptualized as a process occurring *between units;* the sense data of associationism were discrete, sovereign units. The experimental movement left this assumption untouched. To be sure, the actual study of associations has rarely worked with elements; typically, the procedure has been to establish associations between complex contents. But the failure to question the traditional assumption had the following consequences. First, association was always studied between distinct units. Thus, the prevailing assumption perpetuated a uniformity of procedure and discouraged efforts to vary deliberately the relations between the terms that were brought into association. Second, it forestalled questions about the properties of the stimuli that were studied. The study of associations has largely taken the stimulus for granted, concentrating on the connections alone. The result was to maintain the sharp separation between perceptual and associative operations.

The subsequent advances in the field of perception, and in particular the study of unit segregation and formation, have not introduced any substantial modification into this traditional position. The investigator may assume that principles of perceptual organization determine the formation of units. But these perceptual events, he holds, serve only to set the stage for the subsequent, and wholly different, associative operations. Perceptual conditions may determine the identifiability and clarity of the stimulus, but association proper is an activity following its own laws. This is in effect to deny a *direct* influence of perceptual relations on association, while acknowledging that perceptual conditions exert a limiting effect on associative operations. The latter, to be sure, must work on what is perceptually given; perceptual principles decide what the units are, or *what* will be associated. But once these conditions are given, associative operations enter that are autonomous. Thus, the data of perception play a role in current associative thinking analogous to that of the elements in the earlier forms of association doctrine.

The findings of this investigation, which establish a relation between perceptual properties of the stimulus and functional coherence, raise the following questions for the position of associationism. If the phenomena here studied are all instances of one process, it becomes necessary to revise current associative formulations at a decisive point, and to assert that *perceptual relations exert a DIRECT effect on the formation of associations.* This conclusion would come into doubt if it were established that the coherence of terms within a unit is different in process from the coherence between one unit and another, a question to be discussed in the following sections. In that case, it would be necessary to conclude that the traditional study of association has dealt with a restricted question, that it has not encompassed the full range of functional connections, and that the stimulus conditions for association must be specified in perceptual terms.

2. The gestalt studies of perception of the last 50 years have been the basis of a radically different account of the relation of perceptual to associative events. It reverses completely the earlier emphasis, proposing to replace an associative theory of perception with a perceptual theory of associations. This is the intent of the theory formulated by Köhler (1929, 1941). That percepts are organized signifies that the parts are in interaction, or that they mutually alter each other. Now the material of recall preserves, often quite correctly, the organization of earlier experience. It follows that memory traces, or the physiological bases of memory, are the aftereffects of preceding organization. This is to say that the terms associated no longer retain their independence; they are parts of a unified organization. In this formulation, an association is the persistence of a preceding organization, not an event added to the terms brought into association; in short, an association is fundamentally different from a mere connection between terms that remain independent. It also follows that conditions favoring organization in experience favor trace organization, and therefore recall.[19]

The results of this investigation provide, we propose, the most direct evidence for the relation of perceptual organization to coherence in memory. In this sense they furnish confirmation for the theory of Köhler. At the same time, neither of the principal theories of association has dealt directly with the problem of this study, namely, with the distinction between coherence within a unit and the coherence of one unit with another. It remains to be seen what the relation of the present problem and findings is to these positions.

A Further Distinction between Unitary and Nonunitary Coherence?

We have hitherto concentrated on the quantitative differences in memory between two kinds of stimulus conditions. We will now explore the thought that the difference in question is not adequately described in quantitative terms alone, that the two conditions are separated in a more systematic and far-reaching way. Finally, we will consider whether one process of coherence will suffice to account for both sets of data.

The problem can best be introduced by referring to a known fact concerning associative recall, but one that deserves to be better-known. The recall of an association, more particularly the recall of *b* when *a* is given, may fail for two fundamentally different reasons. 1. Recall will fail, at the time of test, if there is failure to recognize the given term *a*. This assertion follows from an assumption we consider compelling, namely, that associative recall presupposes a prior step of recognition (Köhler, 1940). 2. Recall of an association may not occur although the starting term *a* is recognized; in that case there is a failure in the recall *of an association*.

---

[19] The theory of Köhler concerns the effects of organization generally, and is not restricted to perceptual organization. The general formulations are, of course, intended to include perceptual organization.

We now propose to examine the thesis that failure to complete a unitary stimulus has its source exclusively in the first of the two alternatives described above, but failure to complete a nonunitary pair, which often occurs despite recognition of the first member, is due to failure of associative recall. This may be restated positively as follows. When part of a unitary stimulus is recognized, the missing part will be invariably completed; in contrast, one of the members of a nonunitary pair may be recognized without producing completion of the other member. This distinction, if empirically established, might be a criterion of important differences in the underlying processes.

Before proceeding, let us note the conditions that have to be met in order to decide whether failure of recall is due to failure of recognition or of the recall of an association. First, it must be established that the association in question was formed. Second, it is necessary to determine whether recognition of the terms occurs at the time of test. Despite the obviousness of these requirements, there is very little concrete knowledge at present of the contribution of these factors. The investigation of memory has almost totally overlooked the distinction between these sources of failure of recall, a consequence of the neglect of the fundamental difference in process between recognition and recall.

One consequence follows from the stated hypothesis for the data of this investigation. Recall of a unitary stimulus should be *complete*; that is, recall of form and mode should not occur one without the other. (Implicit is the requirement that failure to recall a unitary stimulus should also be complete.) Since the present investigation was not initially designed as a test of the hypothesis, the following re-examination of the data will be somewhat indirect and only suggestive.

1. One source of evidence comes from a re-examination of the matching data of Experiment 8.

Experiment 8 included, in addition to matching, a test of recognition of form. We now selected those forms in the unitary series that were correctly recognized, and computed for each *S* the proportion of cases in which these forms were also correctly matched. The mean of these proportions was 63.8; that is, on the average 64% of recognized forms were correctly matched. It may seem that this result is negative for the hypothesis in question, but this would be an incorrect inference. As we have seen, the criterion put forward requires that the unitary stimulus should have been adequately perceived at the time of learning. If, for example, the *S* initially perceived the form but for some reason neglected the mode, subsequent partial recall would have no bearing on the hypothesis. We lack, however, direct information concerning the perception of modes in this condition. Fortunately, we can obtain an approximation to such a measure from Experiment 6, in which the conditions of learning were identical with those of Experiment 8, but the test was one of recognition of the forms and modes, respectively. The results appear in Table 18.

Proceeding as before, we selected all instances in Experiment 6 of correct recogni-

TABLE 18

RELATION OF RECOGNITION OF FORM TO PAIRING (in %)

	Experiment 6: "Synthetic Pairing"	Experiment 8: Matching: Intentional	Experiment 9: Matching: Incidental
Unitary	70.4	63.8	45.1
Nonunitary	84.0	31.6	19.2

tion of form in the unitary series and computed for each *S* the proportion of instances in which the paired modes were also recognized. The resulting values give us a measure of "synthetic pairing" of recognized forms. The significance of this measure is that it states the limit of actual pairings that can occur under the given experimental conditions. Actual pairing cannot exceed the level of synthetic pairing, which is an index of the joint *perceptibility* of forms and modes that belong to the same stimulus; the former can only be of lower frequency than the latter.

The mean proportion of synthetic pairings in Experiment 6 was 70.4; that is to say, on the average 70% of the recognized forms were accompanied by the recognition of their modes. The difference between this value and 63.8%, the value obtained in Experiment 8, is small and not significant ($p > .10$).

The preceding analysis assumes that the recognition data of Experiments 8 and 6 can be compared despite the differences of conditions at the time of test. (In Experiment 6 new items were interspersed with the old for the test of recognition; not so in Experiment 8. Further, the recognition test of Experiment 8 followed a test of matching.) Despite these differences, the levels of recognition of form were remarkably similar in the two experiments. Therefore, the comparison seems justified as well as the assumption that the two variations produced similar levels of recognition of modes.

When one takes into account the perceptibility of the stimuli in Experiment 8, the completeness of retention of the unitary stimulus is in accordance with the criterion. Nevertheless, a more direct test will be necessary for a definitive conclusion.

2. The data of Experiment 10 are also pertinent. Here we employed a procedure of familiarization preceding the main task, which insured the full identifiability of the stimuli at the time of learning. We also tested for recognition of the original stimulus series. For the purpose of testing the present hypothesis, this procedure may be the most sensitive. Our concern is to establish whether a given unit has been retained as a unit; given this aim, it is desirable that the test reduce the burden of recall proper to a minimum.

The interpretation of the results has to take into account the following detail: the test of recognition was itself a source of interference. We found that the position of the stimuli in the test series affected accuracy of recognition. Accordingly, we examined the level of recognition of stimuli that came *first* in the test series (these stimuli all appeared toward the middle of the learning series, in the fourth position, a position susceptible to interference). The level of accurate recognition of the unitary stimuli was 96% (in contrast to the level of 52% for the nonunitary stimuli). This result is based on one half of the stimuli in the unitary series.

The pertinence of this result derives from the following consideration. The procedure of familiarization made certain that there would be complete recognition of the parts of the unitary stimulus. Under these conditions we find that recognition of the belongingness of the parts is at a maximum, which is in accordance with the hypothesis.

3. Turning to the data of free recall, the hypothesis requires completeness of recall of the unitary stimulus. We have found that such inseparability was by no means always the case: isolated recalls and incorrect pairings did occur with measurable frequency. Before drawing a negative conclusion, the following points need to be considered. (*a*) The criterion in question presupposes, as we saw, that the stimulus was adequately perceived at the time of direct experience. We know that this was not the case throughout, if we may judge by the recognition data (see Experiment 6). (*b*) Further, it was quite possible to register a stimulus simultaneously in a unitary manner and discretely. The modes in particular were often highly familiar; the *S*s were capable of perceiving them as parts of the stimulus and at the same time of recording them separately. Such multiple registration could also account for isolated recall. (*c*)

Finally, uncontrolled guessing could account for certain incorrect pairings. It should be noted in this connection that incorrect pairings occurred infrequently (see especially Experiments 1, 4, and 5 which are most pertinent). The free recall data are thus inconclusive. Failure of recognition, separate registration, and guessing limit their value for the present purpose.

Although the results of the preceding analysis are not definitive, they offer some support for the hypothesis when the necessary conditions are most adequately met. But further investigation, directly oriented to the hypothesis, is clearly needed. Such investigation will require the simultaneous determination of data not ordinarily obtained in experimentation: it will be necessary to establish that learning had occurred, and to relate, at the time of test, the data of recall to those of recognition.

Before concluding, the following remark may be in order concerning the criterion of invariable completion. It would be going too far to assert that the trace of a unitary percept cannot be disrupted under any conditions, that it must always persist—or disappear—as a unit. Our intention is to make a less sweeping assertion, to suggest that a unitary trace will persist as a unit under "normal" conditions, as long as no extreme measures are taken to disturb it. Admittedly, what is to be regarded as normal is left undefined at this point. Despite this shortcoming, and despite the uncertainties surrounding the hypothesis, the criterion directs us toward a valid problem. It asks us to consider what parts, or aspects, of a stimulus distribution are inseparable, or—since the criterion is a severe one—nearly inseparable. It can hardly be said, in the light of present knowledge, that investigation will fail to discover instances that meet the criterion.

Associative and Nonassociative Processes

The evidence for two substantially different kinds of stimulus conditions and effects prompts us to ask whether the respective operations are similar or distinct. Generally, in psychology, we take the joint recall of "two" data as evidence of association. Is it equally clear that the coherence of terms belonging to the same unit is an instance of association? There is the possibility that one uniform process is responsible generally for functional coherence, whether it occurs within a perceptual unit or between units. If so, it would follow that the conditions we have called unitary provided superior conditions *for association*. But we need consider also the alternative that there are distinct kinds of coherence, and that these are a function of the particular unit relations.

This question has not attracted attention mainly because the study of units has been the province of perceptual investigation, while associative studies have limited themselves to the combination of units. It has become evident, though, in the light of this investigation, that the properties of units are pertinent to the general problems of association. Further contributing to the prevailing assumption are the procedures currently employed in investigation, which are best designed to reveal quantitative differences between effects, thus creating the presumption of a uniform process. To speak of association is harmless as long as the term refers to a uniform set of procedures and effects. But as soon as it refers to different conditions and effects, it becomes charged with theoretical implications, and ceases to be merely descriptive. Surely, it would not be right to conclude that the unitary condition is associative simply because the procedure employed in studying it was that of the association experiment.

What grounds may there be for the hypothesis of a nonassociative kind of coherence? We would first mention an observation that is only suggestive. The typical associative situation (as represented in the nonunitary condition) begins with the apprehension of unrelated data that are *subsequently* joined; perhaps it is this transition from an initial to a later state that characterizes the step of association. The situation is importantly different when the data are *from the start* apprehended as members of one unit. In the latter case the relation between the terms evolves directly as a perceptual process, lacking the transition from an initial to a terminal state. We are mindful that this observation is not conclusive; the difference in question could well be functionally continuous.

The following consideration seems to us the most compelling: the unitary condition produced an effect that no other pairing condition achieves. We know of two substantially different association experiments: that which pairs heterogeneous unit terms, and the pairing of units that stand in clear relations of fitting, of which the investigation of Prentice and Asch (1958) is an example. The unitary condition differs from the former of these in ways already described. It is now necessary to call attention to the critical difference between the unitary condition and that produced by related units. In the latter case coherence was also at a high level, but it was a consequence of intrinsic perceptual relations. The unitary condition produced comparable effects when the terms in question were heterogeneous. We know of no stimulus relation other than that internal to a unit that brings about this effect.

We need now consider whether these effects can be referred to a single process, since such a solution would introduce a most desirable orderliness in our thinking. In doing so let us refer to the three conditions of coherence that have been identified: (*a*) the pairing of distinct and heterogeneous units, as in the classical association experiment; (*b*) the pairing of units related by properties such as symmetry, completion, etc. (Prentice & Asch, 1958); and (*c*) the "pairing" of heterogeneous constituents of units. The three cases differ in fundamental respects: (*a*) and (*c*) are alike in heterogeneity but not in coherence; (*b*) and (*c*) are comparable in coherence but not in heterogeneity; while (*a*) and (*b*) differ in both respects. To assert that they are functionally continuous, we believe, requires the following arguments.

One would have first to reverse an earlier assumption and maintain that all of the preceding conditions are unitary, and that they differ solely in degree of unity. More particularly, this interpretation asserts that any pairing, however heterogeneous the terms, forms a unit; that coherence is a function of degree of unity or interdependence among the members of a stimulus distribution; and that the empirical discontinuities are a consequence of an underlying functional continuity. This is the position that Köhler (1929) has taken in a classic discussion, in which he speaks of association as "coherence within the unitary trace of an equally unitary or organized experience. . . ." (p. 282).[20]

To refer all functional coherence to organizational conditions, we believe, is to go in a necessary direction. The formation of the most recalcitrant association is evidence of interaction between the terms in question, and to this extent might also properly be called an instance of unit formation. We are not convinced, however, that this formulation, which was directed to the typical association experiments, fully takes into account the problem which the present investigation poses. The latter requires us to ask whether a unitary relation that is formed spontaneously is functionally identical with the unity established between heterogeneous terms. It is well-known that the properties and effects of interaction can differ considerably; indeed, interaction may also result in segregation. (To judge by the phenomenal evidence, this describes one aspect of the figure–ground relation of Experiment 11.) Thus, the issue is substantially unchanged when one takes into account the organizational character of rote associations.

The following would be a more direct way of deciding between the alternatives. If the stimulus conditions we have contrasted are functionally continuous, this could be demonstrated by finding intermediate stimulus distributions and observing whether the results of coherence are correspondingly intermediate. Such investigation would be timely. We see the possibility of systematically varying for this purpose the degree of fittingness (and of other relations) between paired units. At this time, though, we know of no way to bridge the unitary relation we have here studied with the corresponding dual relation.

---

[20] At another point Köhler (1941) has again asserted that "the fact that a unitary process forms a unitary trace is equivalent to what we call an association. From the present point of view association is therefore simply coherence within the unitary trace of a unitary experience" (p. 493).

If there are distinct kinds of coherence, the resulting memorial processes must be correspondingly different. We will now attempt to derive this difference from the distinction, examined in the preceding section, between recall that invariably follows recognition and recall that may fail despite recognition. For the purpose of discussion we will assume what is still to be proven, namely, that there are specifiable stimulus conditions in which recall follows invariably upon recognition.

The particular statements to be examined are as follows.

1. Completion in memory that invariably follows recognition is an instance, not of recall, as is generally assumed, but of recognition. But completion in memory that may or may not follow upon recognition requires a distinct step of recall.

2. The first kind of completion is nonassociative, the second is an instance of associative recall. This formulation states an independent, although still quite indirect, property of an association: the criterion of an association is that recall of it can fail following upon recognition of the preceding term with which it was paired. This point deserves notice in the light of the observation that the occurrence of an association has hitherto not been defined except in terms of the employment of a particular experimental procedure.

3. It follows that completion in memory of part of a unitary stimulus is an instance of recognition, which is nonassociative, but completion between heterogeneous units is a case of associative recall.

For illustration we compare the following memory results:

1*a*. Recognition of a unit lacking natural subparts.

1*b*. Completion of part of a unit lacking articulated subparts.

2. Completion of a unit with articulated subparts.

3. Completion of heterogeneous unit pairs.

1*a*. An observer inspects a black elliptical contour line on a white ground. At the time of test the same figure appears with a number of others; let us suppose that it is recognized. The occurrence of recognition we interpret in accordance with Köhler's (1940) account of the Höffding function. The perception of the ellipse produces as an aftereffect a trace corresponding in organization to that of the percept. On its later appearance, the elliptical stimulus produces a new percept corresponding to current stimulation. The occurrence of recognition signifies that the latter process makes contact with the trace of the earlier percept, and that the contact is made possible by the similarity between them.

1*b*. The stimulus is as above—an elliptical contour line. At the time of test the observer is shown a portion of the ellipse, with instructions to reproduce the earlier stimulus. For completion to occur, the part-stimulus must be recognized as similar to the one that was seen earlier. That is to say, the present perceptual process, on the basis of similarity, must arouse the trace of the earlier percept. This situation differs in an obvious respect from the one just described earlier. The stimulus is now reduced at the point of test; the trace is contacted by a process that corresponds to it only partially. But the trace is of a unitary form *without* parts; the ellipse possesses a whole-quality that is given directly as a perceptual datum. (A sophisticated observer may of course apprehend the shape in a more analytical way, but we will assume that this is not the rule in immediate perception.) If, as in this case, the trace of the earlier organization is highly unified, it may not be possible to reactivate it without doing so in its entirety. Once trace contact of the kind described has occurred, the entire trace of the old stimulus will be activated, a direct consequence of its unity.

well-known and strong condition of coherence. This effect of meaning has been variously interpreted. Often, it has been referred to familiarity produced by past experience. While this is often the case, one may doubt the adequacy of this interpretation in all cases. A relation of meaning is also an instance of conceptual unit formation; the advantage that meaning introduces is doubtless a consequence of this fact. The question one might raise is whether the contribution of meaning in the setting of the association experiment is at all unique, and whether the advantage it provides is not completely a function of the relation of unity it establishes.

SUMMARY

The traditional study of associations is part of a more inclusive problem, one that concerns the relation between the objective structure of the stimulus and the coherence of its parts, as evident in perception and memory. The classical association experiment has restricted observation to a limited phase of this problem. It has concentrated on the coherence established between pairs of heterogeneous units. Theory and investigation in this area have not considered the formally analogous operations within well-structured units. This study was concerned with the effect of the perceptual relation of unity between the parts of a stimulus distribution on the capacity to retain them jointly.

Employing the technique of the classical association experiment, and visual forms as stimuli, the perceptual relations between the parts of a stimulus distribution were systematically varied. The following conditions were studied: the relation of constitution or composition, the relation of figure and ground, the relation of inclusion, and the relation of contiguity between pairs of heterogeneous units. The contents or terms composing the several relations were either identical or closely similar. The relations in question were phenomenally describable as either unitary or nonunitary.

Coherence of recall, or the ability to recall jointly the parts of a stimulus distribution, was substantially superior when the data belonged to the same perceptual unit. Errors in recalling "what went with what" were regularly more frequent when the parts of a stimulus distribution belonged to perceptually distinct units. These differences persisted with different tests of recall, and also when the test was that of recognition, thus establishing that the effect occurred at the point of direct experience.

The results demonstrate a direct relation between the unit properties of the stimulus and coherence. Given a difference of the unit relation between two stimulus distributions, the direction of difference between them in coherence of recall can be predicted.

The question was considered whether coherence within and between units may differ in process. The hypothesis was examined that the spread of excitation from one unitary trace to another requires a step not present in the spread of excitation within a trace. A test of the hypothesis was proposed, which rests on the differential role of recognition in the unitary and nonunitary instances. A re-examination of the empirical data offered a degree of support for the hypothesis, but a definitive conclusion awaits further investigation that meets strictly the experimental conditions required by the tests.

The consequences of the preceding hypothesis for a theory of associations were examined, as was also its relation to current theories of associations.

The relation of this study to a range of questions in general psychology was discussed.

REFERENCES

GOTTSCHALDT, K. Über den Einfluss der Erfahrung auf die Wahrnehmung von Figuren. *Psychol. Forsch.*, 1926, **8**, 261–317.

KÖHLER, W. *Gestalt psychology.* New York: Liveright, 1929.

KÖHLER, W. *Dynamics in psychology.* New York: Liveright, 1940.

KÖHLER, W. On the nature of association. *Proc. Amer. Phil. Soc.*, 1941, **84**, 489–502.

PRENTICE, W. C. H., & ASCH, S. E. Paired association with related and unrelated pairs of nonsense-figures. *Amer. J. Psychol.*, 1958, **71**, 247–254.

THORNDIKE, E. L. *Fundamentals of learning.* New York: Bureau of Publications, Teachers College, 1932.

(Received February 6, 1959)

APPENDIX

SEPARABILITY OF FORM AND MODE

Units differ in degree of coherence. There is also reason to believe that units differ with respect to the range of conditions under which they will remain coherent. Indeed, resistance to disturbing conditions may be an index of degree of unity. The following experiments were an attempt to explore the unity of the form–mode relation by means of a new procedure.

In the preceding investigation the *S*s were uniformly directed to note both the form and mode of each stimulus. We now undertook, by means of altered instructions, to see to what extent it is possible to direct the *S* exclusively to the form or mode alone of the unitary stimulus. The degree to which this separation can be achieved may serve as an index of lack of unity. In short, we take *separability* of the properties of a unitary stimulus as a criterion of unity.

Experiment 15: Separability of Form from Mode

The unitary stimuli of Experiment 5 were the materials of this experiment. The procedure was that of Experiment 5, except for the following modification. The *S*s were instructed that the task was to recall only the *forms*. The instructions stated in part: "Your task is to recall the larger shape alone. You may be interested to know that we are comparing the recall of the larger forms under this condition with one in which the contour line is always drawn continuously (an illustrative card with the same form in continuous contour and with another modal contour was here shown). It just so happens that you have been assigned to the group with different contour lines. After I have shown you the cards once, I will ask you to draw as many of the forms as you can remember, using a continuous line, such as this (pointing to the illustrative continuous sample)." This explanation provided an effective rationale.

The test, which came after 3 min., was one of free recall. At this point, however, the *S*s were instructed to reproduce, whenever possible, the modes as well as the forms as they were initially given.

The design of this experiment thus combines an intentional with an incidental procedure. It directs the *S*s to focus on one aspect of the stimulus—the form—and to disregard the other; at the same time the perception of the one must carry with it the registration of the other. In this respect the technique is similar to that of the Külpe abstraction experiment.

Experiment 16: Separabiliy of Mode from Form

The stimuli and the procedure were identical with those above, except that the *S*s were instructed to recall the *modes* only. Thus, this condition was incidental for the form, and intentional for the modes.

The results of both conditions appear in Table 19.

TABLE 19

RECALL SCORES

Experiments 15 and 16

	N		1 +(FM)	2 −(FM)	3 F	4 M	*M*
Experiment 15	12	*M*	3.2	1.2	3.8	0.1	8.3
		%	38.4	14.1	46.5	1.0	
Experiment 16	17						
a. Totally obedient *S*s	5	*M*	2.8	1.2	0.2	5.2	9.4
		%	29.8	12.8	2.1	55.3	
b. Totally disobedient *S*s	3	*M*	6.7	1.3	0.7	1.0	9.7
		%	68.9	13.8	6.9	10.3	
c. Intermediately obedient *S*s	9	*M*	6.2	0.7	0.8	2.9	10.6
		%	60.2	6.5	5.4	27.9	

1. We will first examine the results of Experiment 15.

Perhaps the most noteworthy finding is that the *S*s were capable of complying with the instructions. Of a total of 14 *S*s, 12 were completely obedient to the instructions, judged by their answers to questions, at the conclusion of the experiment, concerning the procedure they adopted during learning.

The quantitative results, which appear in Table 19, provide convincing evidence of the effectiveness of the instructions. Most pertinent is the comparison of form and mode recall. The mean recall of forms was 6.0; of modes, 2.3. Each *S* recalled more forms than modes. Equally telling are the data for isolated recalls: of 47 isolated recalls, 46 were of forms.

As further evidence of the effect of this condition, we may cite the following. At the conclusion of the experiment the *S*s were shown the modes, each drawn as a straight line, and tested for recognition. (The decision to proceed in this way was made when the experiment was already in progress, and was applied to only 8 *S*s.) Although this was not an adequately controlled test of recognition, the results are instructive. The mean number of modes recognized was 5.3, a level considerably below that obtained in Experiment 6. Individual recognition scores ranged from 2 to 8.

In the light of these findings, it is not surprising that the level of joint recall was low. In absolute terms the level was one fourth of that in the U condition of Experiment 5, and less than one half of that in the non-U condition. Correct pairings were also depressed proportionately, when compared with the level of the non-U series of Experiment 5.

2. The majority of *S*s of Experiment 16 found it far more difficult to obey the instructions. They reported that the form remained dominant, and that they involuntarily referred to it when attempting to commit the modes to memory. Compliance with the instructions was again judged on the basis of the *S*s' replies to questions put at the conclusion of the experiment concerning their procedure during learning, and without knowledge of their recall performance. Out of a total of 17 *S*s, only 5 succeeded in obeying the instructions; 3 were completely disobedient; and 9 were judged to be intermediate.

In Table 19 we categorized the recall data in accordance with this classification. Despite the small numbers of cases, the results fall into a clear pattern.

a. Fully obedient *S*s performed in a manner that completely corresponds to the performance of nearly all *S*s in Experiment 15. There was a striking preponderance of mode over form recall, the respective means being 7.2 and 2.2. Out of 27 isolated recalls, 26 were recalls of mode. All 5 *S*s showed the same superiority of form over mode recall.

Correct pairings were markedly depressed; there were only 7 such instances. Even this figure is inflated, since all cases of joint recall were of the first two stimuli in the learning series, that is, of those appearing during the phase of least crowding.

Thus, separability of form and mode was achieved in this, as in the preceding experiment, although by fewer *S*s.

b. The three totally disobedient *S*s differed strikingly from the preceding group. Recall of forms and modes was now identical. Most noteworthy was the drop of isolated recalls. In all respects these *S*s resembled closely the U *S*s of Experiment 5; this was so especially with respect to the frequencies of correct and incorrect pairings. Further, correct pairings were no longer confined to the early stimuli in the series.

c. The intermediate *S*s fell between the two preceding groups. There was a preponderance of mode recall; the latter accounted for the 61.3% of all recalls. Pairing was close to the U variation of Experiment 5.

We tested all *S*s for recognition of form. These were shown at the conclusion of the free recall test, drawn in continuous contour. The mean recognition score was 5.9, again considerably below the level obtained in Experiment 6. Individual recognition scores ranged between 2 and 10.

Complete separability was not achieved in either experiment. But a substantial degree of separability occurred under both sets of instructions.[21] Also, it was easier to focus on the form than on the mode, a result in accordance with earlier evidence that recall of form predominated in the unitary condition (see pp. 8, 11).

A full interpretation of these findings must await further investigation of the relative ease of separation of unitary stimuli of other kinds.[22] It would be of particular interest to identify those part-properties of stimuli that cannot be divorced from each other in perception and recall.

Under the conditions here described, one property of a stimulus was made focal and another peripheral; at the same time the perceptual situation was such that the ignored property could not but be registered. It was hardly possible to inspect the form without registering the modes as a visual datum, and conversely. That the ignored property suffered in recall, and even more important, in recognition, is evidence of the importance of attention, as distinct from sheer stimulation, in learning.

---

[21] The procedure of testing for recall probably underestimates the actual level of coherence. This in no way weakens the weight of the comparison between the present data and those obtained under the fully intentional conditions of Experiment 5, which also employed the test of free recall.

[22] We have not done corresponding separability experiments with the nonunitary series. Such investigation might be appropriate as part of a systematic inquiry, but the results obtained earlier with the nonunitary series justify the surmise that such a procedure would probably be extremely effective.

The following caution must be noted in interpreting the preceding results. Separation of form and mode was achieved only as a consequence of a special attitude induced by instructions to the *S*. It would be unwarranted, and probably misleading, to infer that form and mode were equally separable under the conditions of the main experiments of this investigation, and that this separability was counteracted by an equally special attitude in the opposed direction. It is more likely that we have here tested separability under extreme conditions, and that neutral conditions would demonstrate a considerable coherence of form and mode. These comments make clear the need for study of the coherence of the form–mode relation, as of other units, under "neutral" conditions, such as will avoid directing the *S* either to the entire stimulus distribution or to any aspect of it. Such a procedure should throw light on the modes of apprehension of stimulus relations under relatively spontaneous conditions.

Vol. 74, No. 4 Whole No. 491, 1960

Psychological Monographs: General and Applied

MEANING AND FAMILIARITY IN ASSOCIATIVE LEARNING[1]

WILLIAM EPSTEIN
University of Kansas

IRVIN ROCK
Yeshiva University

AND CARL B. ZUCKERMAN
Brooklyn College

It is taken for granted that meaningful material is learned more easily than nonmeaningful material. McGeoch and Irion (1952) write that "a high positive correlation between meaningfulness of material and rate of learning holds under a very wide range of conditions" (p. 479). Thus, when subjects are given the task of learning a list of meaningful items (e.g., adjectives) and a list of nonsense syllables of identical length, the adjective list is learned to criterion in fewer trials. Also, structured sequences of words (e.g., sentences) are easier to learn than lists of randomly selected words.

Demonstrations of this nature together with other related experimental findings have led to the general acceptance of the facilitating role of meaningfulness. They have not, however, produced any unanimity in accounting for this effect of meaning. Five principal explanations which have been advanced are:

1. *Associationist explanation.* Meaningfulness as a determinant of learning gains its status as a consequence of positive transfer. Meaningful material allows the learner to draw on a vast store of previously established associations. As a result, the present process of learning has only to strengthen some particular bond between the items which was already in existence. Pairs of nonsense syllables, on the other hand, do not enjoy the advantages of previously formed associations; since they require the formation of new associational bonds, they are more difficult to learn.

Recently, this point of view has been restated in a somewhat modified way. For example, Mandler and Huttenlocher (1956) state that the learning of new associations is "mediated by previously learned associations to the stimulus- and response-terms" (p. 424). Therefore, "the more such associations are elicited by the stimulus- and response-members in a paired-associate task, the greater will be the probability that a particular combination of a stimulus-association and a response-association can be found to provide a mediating bridge or a 'meaningful' link between the two" (p. 424). Mandler and Huttenlocher asked subjects to associate to each of a number of nonsense syllables for an interval of 30 seconds. The syllables were then ranked according to associative frequency (i.e., the number of associations evoked). Subjects were then given the task of learning a list of eight pairs of syllables, each pair having a different level of associative frequency. A significant correlation was obtained, "between speed of acquisition of a nonsense-syllable pair and the mean associative frequency previously established for its two components" (p. 428).

Presumably, the same reasoning could be applied to explain the facilitative effect of meaning. A meaningful word would evoke a greater number of associations and, therefore, pairs of such words would be more likely to have "mediating bridges."

2. *Organizational explanation.* A second point of view, represented by Köhler

---

[1] This investigation was conducted when the first two authors were at the New School for Social Research and was supported in part by Research Grant M-2082, National Institute of Mental Health, Public Health Service.

(1929), contests the associationist explanation "mainly because the same words have occurred thousands of times in other much more regular connections, which ought absolutely to inhibit the assumed associations in this particular case" (p. 287). This inhibition is to be expected particularly where one takes pains to pair words which would not have been experienced together often, if at all, in the past. Köhler argues that the emphasis on association is misplaced and that the true explanation must be made in terms of the possibilities for organizing the learning material. That is, if while reading the pair of nouns, pen—girl, the learner visualizes a girl writing with a pen, then the learner will have "experienced some well organized . . . whole,[2] and it may be that here learning is so easy because organizing is so much easier in this case than in that of indifferent nonsense syllables" (Köhler, 1929, pp. 287–288). If we determine which material is more amenable to organization and in turn determine where association is strongest, we shall find them coincident. Thus, "where organization is naturally strong we have spontaneous association; where there is practically no organization association does not occur until some organization is created intentionally" (Köhler, 1929, p. 290). Köhler points out that if the noun pairs, lake—sugar, pencil—gasoline, etc. are presented to a group of subjects, they will be learned more easily than an equal number of nonsense syllable pairs. Certainly, Köhler argues, no one can say that between such categorically disparate items as pencil and gasoline there exist old associations which merely require some degree of repetition or exercise for success in the learning task. If learning is facilitated under these conditions it can be only as a result of organization. It should be noted that Köhler's arguments do not apply to the more recent versions of the associationistic view, e.g., Mandler and Huttenlocher.

3. *Familiarity explanation.* This approach, which is similar to Robinson's (1932) "law of acquaintance" would view the theories described above as superfluous. The crucial factor is the familiarity of the individual items used in the learning task. Meaningful material is also familiar; therefore, it would be easier to master a list of pairs of meaningful items than nonsense items. When adjectives are used the learner does not have to be concerned with mastering the individual items; these are already familiar and thus easily distinguishable. He can concentrate mainly on forming an association between the members of a pair. Where nonsense material is used, the learner must first learn the individual unfamiliar items and, in addition, form an association between the items. The additional task would make nonsense material harder to learn.

There is some experimental evidence bearing on this argument. Waters (1939) reported that prefamiliarizing nonsense syllables did not facilitate learning. He demonstrated that varying amounts of prior reading and reciting of syllables did not significantly facilitate learning them in a new order. In his experiment three conditions were compared:

Condition I. The subjects first read aloud the list to be learned three times. They then proceeded to the formal learning of the list. The serial order for learning was the same as for initial reading.

Condition II. The same as Condition I except that the order of syllables for learning was different from that during the initial reading.

Condition III. Learning without any prior reading of the list.

Waters considered the comparison between Conditions II and III to be the crucial test for the effect of familiarization. He found no difference in the number of trials to learn the lists under the two procedures.

Sheffield, however, did not confirm Waters' conclusions. In Sheffield's (1946) study, syllables were prefamiliarized by a procedure which did not involve the formation of associations between syllables. In the prefamiliarization session the subject saw the word or syllable, vocalized it, then tried to find it quickly in a list of three-letter words. When the word was found, it was marked and vocalized again. These items were then paired and compared with pairs of new items. The prefamiliarized pairs were learned more easily than the new pairs.

---

[2] We shall refer to such organized wholes as "conceptual units."

A third study on the role of familiarity has been reported by Hovland and Kurtz (1952). These authors suggest that Waters failed to demonstrate an effect of prior familiarization because his technique required the unlearning of associations which were formed during learning. In their experiment, 12 subjects learned lists of 6, 12, and 24 nonsense syllables, both with and without prior familiarization of the items. The prefamiliarization technique consisted in having the subjects study the nonsense syllables which were presented in constantly changing order until they knew the syllables well enough to recognize them and also to supply the missing letters when parts of the syllables were presented. The effect of prior familiarity with the syllables was to reduce the number of trials to mastery by a moderately constant amount for all lengths of lists.

While familiarity (in the sense described above) has been shown to play a role in associative learning, the evidence is not sufficient to prove that the superiority of meaningful material is entirely due to this factor.

Explanations closely allied to the familiarity and association approaches have appeared in recent years. These stress the role of linguistic structure and patternings rather than the familiarity of the items or the mediation by specific associations:

4. *Transitional (or sequential) probability.* Words or letters in speech and writing do not appear in random sequences but follow one another in predetermined orders. Transitional analysis investigates the dependent probabilities of sequences of linguistic units:

Given the unit "x," what is the probability that it will be followed by unit "y"? Some sequences never appear in English while others vary in frequency of occurrence. The letter "q," for example, is never followed by any letter other than "u"; there is a greater likelihood that "e," as the first letter of a word, will be followed by "n" or "s" than by "y." The word "is" or "says" will more likely occur after "he" than will "electron." Higher order probabilities can be obtained by posing the question: given the sequence of units XY or XYZ, what is the probability that a particular unit will be the next to appear?

Miller and Selfridge (1950) argue that sequences of high transitional probability will be easier to learn than those of low frequency. To obtain evidence for this hypothesis, they constructed word sequences with different orders of approximation to standard English. For example, a zero-order was derived by selecting words at random from a dictionary. For a first-order approximation, a word is chosen if it frequently follows one word before it; for a second-order, two words before it, etc. The Shannon technique was used to generate the various degrees of approximation. The subjects were given the task of learning 10-, 20-, 30- and 50-word passages ranging in sequential probability from a zero-order to text passages. The results based on an immediate recall test are in accordance with the hypothesis that higher-order sequences are learned more readily than are lower ones.

An important finding in this experiment is that above the fifth-order there is very little improvement in recall for all sequence lengths. Some passages of low order were learned as well as text passages. For example, the third-order 20-word nonsense sequence—"family was large dark animal came roaring down the middle of my friends love books passionately every kiss is fine"—was learned as well as a "sensible" passage of 20 words taken from text.

Miller and Selfridge (1950) conclude from this result that it is not meaning per se which facilitates learning but the presence of sequences frequently encountered in the past:

> Thus there are kinds of nonsense that are as easy to recall as are meaningful passages. The significant distinction is not to be drawn between

> meaning and nonsense but between materials that utilize previous learning and permit positive transfer and materials that do not. If the nonsense preserves the short range associations of the English language that are so familiar to us, the nonsense is easy to learn (p. 183).

Word pairs or sequences which occur together frequently in speech and writing have built up strong associations and are therefore easier to learn when presented in an experiment than are nonsense syllables or random assortments of words.

It is apparent that the concept of transitional probability as applied to the problem of associative learning is essentially a restatement of the first approach considered above.

5. *Grammatical habits.* Almost all of the verbal arrangements we experience and produce are ordered according to the rules of syntax. Nevertheless, a sentence, for example, cannot be defined simply as a highly probable sequence of words. Linguists distinguish between syntactical order and dependent probability computations.

"One's ability to produce and recognize grammatical utterances is not based on notions of statistical approximation and the like" (Chomsky, 1957, p. 16). Again, "If we rank the sequences of a given length in order of statistical approximation to English, we will find both grammatical and ungrammatical sequences scattered throughout the list; there appears to be no particular relation between order of approximation and grammaticalness" (Chomsky, 1957, p. 17). Grammatical structure can be easily recognized even in an arrangement of nonsense syllables, e.g., "The vapy koobs desaked the citar molently" (Fries, 1952, p. 111).

It can be assumed that the great frequency of grammatical sequences in our experience has given rise to verbal habits or learning sets. These habits would facilitate the learning of new material arranged in accordance with grammatical structure; in this case, there will be positive transfer from previous learning of material with the same structure. Ungrammatical or random sequences of words would be more difficult to learn.

Some evidence for this view is cited by Osgood (1957):

> . . . one of my students, Mr. Albert Swanson, compared the ease of learning nonsense sequences that retained the structure of the English sentences from which they were derived, for example,
>
> The maff vlems oothly um the glox nerfs,
>
> with matched materials in which the grammatical cues had been eliminated, for example,
>
> maff vlem ooth um glox nerf.
>
> The structured series were learned more easily than random lists (p. 88).

We have reviewed several explanations for the superiority of meaningful material in learning. Four of these account for the effect in terms of transfer from previous learning of one kind or another—of specific items, associations, verbal sequences, or grammatical habits. Köhler's explanation is the only one based on the referential aspect of language. As we have seen, there is some evidence that familiarity factors are important. But this kind of evidence does not necessarily rule out the role of meaning.

In the experiments to be reported in this paper we attempt to distinguish meaning from familiarity and to investigate the separate contributions of these factors in paired-associate learning.

We know of only one study—by Sheffield (1946)—which deals with this problem. As stated above, Sheffield found that familiarized items are easier to associate than are novel ones. He also raises the question as to whether meaning makes an additional contribution in the learning process. Sheffield defines meaning in S–R terms as the perceptual response which is aroused by a stimulus pattern; this perceptual response is implicit and has been connected to the stimulus pattern by past learning. According to this definition, all familiar stimuli have meaning based on past experience and even novel stimuli may have some meaning based on stimulus generalization. Sheffield attempted to make an item meaningful by attaching a distinctive perceptual response to a word or syllable ("symbolic training"). For example, when the subject saw item X, he was to circle it with a green crayon; item Y was to be blacked out with a pencil, etc. In familiarity training the response was the same to all stimuli used.

The paired-associate learning task was designed so that prefamiliarized material and material which had been made "meaningful" by symbolic training could be compared with novel items. The results show that familiarity and meaning were equally effective. "Symbolic training had no effect other than the correlated familiarization" (Sheffield, 1946, p. 194).

More recently, other stimulus–response definitions of meaning (Osgood, 1953; Noble, 1952) have been proposed. In the present study we are not concerned with the problem of the functional nature of meaning—whether it consists of autonomic responses or cortical events. For our purposes, the dictionary definition of meaning, as that which is named or referred to by a word, seems adequate. The sound CAT is meaningful when used by the speaker and comprehended by the listener to refer to a particular object or class of objects. Words may also refer to events, activities, abstract concepts, etc.

Of course, a word must be familiar in order to be meaningful; a novel word (in isolation) could have no meaning. Nevertheless, familiarity in itself may not give rise to meaning. The syllable "GYX" may become quite well-known with repetition, yet when heard or read it does not arouse any experience beyond itself—no object, concept, or activity is referred to or named. It is therefore meaningless.

Some classes of familiar linguistic units possess structural meanings (Fries, 1952), i.e., they are meaningful only as parts of a sentence. In isolation such words as "who," "the," "for," etc., do not refer to or identify any kind of experience. Words like "cat," "table," "run," "talk," "war," "pretty," on the other hand, do not have to be given in a larger verbal context to have meaning. Presented by themselves, these words will call up specific mental content. One has no difficulty in defining any one of them; "the," however, can be defined only as a member of a particular class of words, or as having a particular grammatical function.

We shall, therefore, use the term "meaningful" to refer to types of words such as nouns, verbs, and adjectives, for which lexical meanings can be readily elicited when they are presented by themselves. The term "familiar" will be applied to nonsense syllables which have been experienced previously and also to types of words such as prepositions and conjunctions. (It is perhaps preferable to say that these latter words have meaning only in the limited sense described above than to call them entirely meaningless.)

The hypotheses which guided the experimentation to be reported were:

1. Familiarity by itself will facilitate associative learning.

2. Meaning will enhance learning even more because it makes possible organization or conceptual unit formation as described by Köhler.

PLAN OF RESEARCH

To obtain evidence for these hypotheses the following studies were carried out:

1. The effect of familiarity. For this purpose, prefamiliarized nonsense syllables were compared with novel ones (Experiment I).

2. The effect of meaningfulness. Meaningful words (nouns) were compared with equally familiar words of limited meaning (conjunctions and prepositions) and nonsense syllables (Experiments II and III).

3. The explanation of the superiority of conjunctions and prepositions over nonsense syllables found in Experiments II and III. To see whether this result is due to familiarity or to the limited meaning of the former items, conjunctions were compared with familiar nonsense items (Experiment IV).

4. The role of transitional probability. To see whether the superiority of meaningful words over merely familiar ones (Experiment II and III) can be attributed to this factor, we compared noun pairs with pairs of words which possessed high transitional probability (Experiment V).

5. The role of organization or conceptual unit formation. Pairs of pictures representing specific objects were compared with concrete noun, abstract noun, and verb pairs

(Experiment VI). Further, pairs of nouns joined by a "sensible" connecting word were compared with pairs having an irrelevant connective (Experiment VIIa) and, finally, pairs of pictures of objects which were arranged to form a unit were compared with picture pairs in which the items remained separated as in the traditional paired-associates method (Experiment VIIb).

Experiment I

In the first experiment we were interested in demonstrating the facilitative effect of familiarity with material made familiar during the experimental period.

Method and Materials

The general procedure consisted in prefamiliarizing nonsense syllables which were later formed into pairs to be learned as paired associates. Learning of the prefamiliarized nonsense syllable list of six pairs was compared with the learning by the same subjects (*S*s) of a list of six pairs of new nonsense pairs under identical learning conditions. All of the nonsense syllables were chosen from Glaze's (1928) list of syllables at the 80% level of associative value.

The prefamiliarization technique employed in this experiment was designed to overcome certain difficulties present in methods previously used by other investigators. The *S*s were required to learn 12 short lists, each consisting of four numbers and one nonsense syllable. Each item in the list was printed individually on a blank card. *S*s read the items aloud and received two trials with every list; a test of total recall followed each trial. In almost all cases the *S* was able to recall the syllable perfectly after the first trial. None of the *S*s suspected the purpose of this procedure. The instructions were as follows:

> This is an experiment to determine how well people learn certain material. I have made up 12 short lists each consisting of 5 items, 4 numbers, and a nonsense syllable (explained if necessary). I want you to try to learn these lists one by one in the following way. . . . At no time will I ask you to recall a list which you have already passed and been tested for, so don't try to keep them all in mind or your task will become impossible.

This procedure differs from previous techniques in two respects. First, the method ensures familiarization by inducing an intention to learn as contrasted with a purposeless repetition of the material (cf. Waters). Moreover, the material is not presented in serial order. The technique of spacing reduces to near zero the possibility that any obtained effect of prior familiarization can be attributed to the existence of remote associations between the syllables rather than to familiarity as such.

The 12 syllables thus familiarized were then paired, thereby constituting one learning list of six pairs. Every effort was made to reduce to a minimum any factors which might cause one pair to be easier than any other, e.g., both members of a pair having identical initial consonants, etc.

The pairs were printed on blank 3 × 5 cards in capital letters. The stimulus item was printed on the bottom left-hand corner of the other side of the card. During learning, each pair was presented for 5 seconds (as measured by metronome) with a 2-second interval between each pair. During the test trials the stimulus item was presented alone for a 5-second interval with a 2-second interval between presentations.

Subjects. Forty-eight college students participated in this experiment. Thirty-two *S*s served in the main experiment and 16 *S*s took part in a supplementary experiment.

Experiment I. The *S*s in the main experiment were divided into two groups. The 16 *S*s in Group 1 were treated in the following way: (*a*) Prefamiliarization session as described above. (*b*) Two learning trials with the list of now familiar items. (*c*) Test after an interval of 1 min. following the second learning trial. (*d*) Two learning trials with new pairs of syllables. (*e*) Test after a 1 min. interval following the second trial. For 8 *S*s of Group 1 the prefamiliarized syllables consisted of those syllables which had been new for the other 8 *S*s. Thus, each item served both as familiar and as new for different *S*s. In Table 1 we refer to this fact by recording the data under two headings, List A and List B. Thus syllable DEV, which was new in List A, was familiar in List B. The results for Group 1 are found in Table 1. The mean number of prefamiliarized pairs learned was 3.1, and the mean for the new pairs, 2.1.

The same procedure was employed for the 16 *S*s in Group 2 except that *S*s in this group always learned and were tested on the new material first and then proceeded to prefamiliarization, followed by learning of the familiar list and test. The results for this group are also presented in Table 1. The mean number of new pairs learned was 2.6, and 3.8 for the familiar pairs.

This experiment was then repeated for the following reason: the new pairs in

List B of the main experiment fared pretty well (particularly in Group 1 and to a much lesser degree in Group 2), with the result that the differences between familiar and new in List B are not as impressive as the differences obtained with List A. For this reason, the 12 syllables making up these six new pairs were interchanged to form different pairs which then constituted the new (nonfamiliar) list for the 16 *S*s participating in this experiment. The syllables to be prefamiliarized were the same syllables which were familiarized in List B of the main experiment. Group 3 consisted of eight *S*s who learned the nonfamiliar list first and then the familiar pairs. These *S*s retained a mean number of 2.9 new pairs and 3.9 familiar pairs. Group 4 consisted of eight new *S*s who learned the familiar list first. The mean number learned was 2.5 for the familiar and 1.4 for the nonfamiliar. Consideration of the frequencies in Table 1 shows that reordering the pairs had the effect of bringing out the difference between the familiar and nonfamiliar syllables. This confirms our conjecture that the factor responsible for the lack of significant difference for this material as given in the original List B is not the ineffectiveness of familiarization but rather the very high level of learning for the new pairs by the *S*s in this group.

Inspection of the last column of Table 1 reveals that the difference between the means for the prefamiliarized and new-syllable pairs are almost identical in the various experimental conditions (1.0, 1.2, 1.0, 1.1), and all are in the expected direc-

TABLE 1

Number of Prefamiliarized and New Syllable-Pairs Learned in Experiment I

Experiment I				
	Group 1			
	List A	List B	Lists A & B	Comb. Mean
Familiar	26	24	50	3.1
New	13	21	34	2.1
N	8	8	16	
	Group 2			
	List A	List B	Lists A & B	Comb. Mean
New	19	23	42	2.6
Familiar	31	30	61	3.8
N	8	8	16	

	Group 3		Group 4	
	Order New-Familiar	Mean	Order Familiar-New	Mean
New	23	2.9	11	1.4
Familiar	31	3.9	20	2.5
N	8		8	

TABLE 2

COMBINED RESULTS FOR EXPERIMENT I

Material	Number of Pairs Learned	Mean Number of Pairs Learned	*SD*
New	110	2.4	1.36
Familiar	162	3.4	1.35
N	48		

Note.—$\bar{x}_n - \bar{x}_f$ is significant at the .01 level; $t = 5.3$.

tion. In Table 2 we have combined the data for all conditions. When this is done, the mean number of nonfamiliar pairs learned is 2.4 and the mean number of familiar pairs is 3.4 for 48 *S*s. The standard deviation is 1.36 and 1.35 for nonfamiliar and familiar pairs, in that order. The difference between the means in favor of familiar pairs is significant at better than the .01 level ($t = 5.3$; $df = 47$).

Experiment II

In this experiment, we used items which are equally familiar but which differ in meaningfulness—we compared nouns with conjunctions and prepositions. If the two types of words were learned equally well, the meaning hypothesis would be invalidated; if, however, an advantage were to be shown by the nouns, then such results would be attributable to the factor of meaning. Nonsense syllables were also used, to be compared with the conjunctions.

Materials and Procedure

Two series composed of three different types of material were used. Each series consisted of three lists of six pairs of items: a list of conjunction and preposition pairs composed of three conjunction pairs and three preposition pairs, a list of six noun pairs, and a list of six nonsense syllable pairs. Two series were used in order to allow for a larger sampling of verbal material and thus reduce the possibility that our results might be attributable to the particular words chosen. An attempt was made to use noun pairs consisting of members not ordinarily paired in everyday use. A check of the Thorndike-Lorge (1944) word count showed that, in general, the nouns, conjunctions, and prepositions we used occur with equal frequency. The nonsense syllables were chosen from Glaze's list of 80% syllables. All of the lists are reproduced in Table 4.

Each pair was printed on a blank white index card. The individual left-hand test items were printed on the back of these cards. Each list was presented to the *S* for one trial, and each pair in this list was presented for 5 seconds as indicated by a metronome. The *S*s were asked to read the pairs aloud (to pronounce, not to spell, the syllables) and to try to associate the items of each pair with one another in preparation for a test of their memory which would follow soon after learning. The *S*s were informed that the sequence was not to be learned since the order of the pairs would be randomized for the test presentation. After one learning trial, the *S*s were asked to work on a paper and pencil maze. This interpolated activity lasted for 1 min. The *S* was then tested by being asked to recall the right-hand member of the pair upon presentation of the left-hand item. This procedure was followed with all three lists in each series. The order of presentation during the test was always randomized by shuffling the cards to a new order. The order in which the different lists (materials) was presented was systematically prerandomized so that the different types of materials followed and preceded each other an equal number of times. Also the order of pairs within each list was changed for each *S*. Each test item was presented for a 4-second interval.

Subjects. Thirty-six college students participated in this experiment. One group of 18 *S*s learned Series I; a second Group of 18 learned Series II. The *S*s were assigned in alternating order to one or the other Series.

Results

The results for this experiment are presented in Tables 3 and 4. For Series I the mean number of pairs learned was 2.7, 1.8,

TABLE 3

NUMBER OF PAIRS LEARNED IN EXPERIMENT II

	Series I	Series II	Series I & II	Combined Mean	*SD*
Nouns	49	55	104	2.9	1.79
Conjunctions & Prepositions	35	31	66	1.8	.97
Nonsense Syllables	12	16	28	.8	.87
N	18	18	36		

Note.—The difference between the mean number of noun pairs retained and the mean number of conjunctions and prepositions was significant at the .01 level ($t = 3.9$). The difference between the mean number of conjunctions and prepositions retained and the mean number of nonsense syllables retained was significant at .001 level ($t = 6.7$).

TABLE 4

ITEM ANALYSIS OF RESULTS IN EXPERIMENTS II AND III

Number of *S*s Who Responded Correctly to Each Pair

Material Series I	Experiments[a]		Material Series II	Experiments[a]	
	II	III		II	III
car—leaf	9	11	boat—wall	10	13
sink—park	9	11	match—rent	9	15
fish—rug	8	12	mat—card	8	12
air—chin	10	13	slip—dog	8	13
wood—shirt	7	14	suit—key	10	12
knee—mask	7	12	pen—sand	10	17
hence—or	6	10	still—if	5	9
also—but	8	9	now—and	6	10
then—yet	6	9	how—nor	4	12
at—from	5	10	in—by	6	10
off—out	5	10	to—for	6	10
down—of	5	8	with—up	4	11
vax—neb	1	6	gos-bif	1	9
muc—yeg	2	9	fij—koz	3	6
het—tas	2	5	jor—gan	4	8
fer—wid	4	8	ril—dut	3	7
mot—daw	3	7	pir—yan	2	8
lav—fow	0	5	dev—ruc	3	7

[a] 18 *Ss* served in each condition.

and .7 for the nouns, conjunctions and prepositions, and nonsense syllables, in that order; for Series II the mean number of pairs learned was 3.1, 1.7, and .9. When the data for Series I and II are combined, the mean number of pairs of nouns, conjunctions and prepositions, and nonsense syllables was 2.9, 1.8, and .8, respectively. The standard deviations were 1.79, .97, and .87, also in the same order.

An item analysis of these results is presented in Table 4. All the differences were highly significant (using a one-tailed test for paired values, nouns vs. conjunction–prepositons, $t = 3.9$, significant at .01 level with 35 *df;* conjunction-prepositions vs. nonsense syllables, $t = 6.7$, significant at .001 level with 35 *df*). An analysis of individual *S* performance revealed that 24 of our total group of 36 learned more noun pairs than conjunction-preposition pairs, 5 *S*s an equal number of both kinds of material, and 7 *S*s more conjunction and preposition pairs. Twenty-six *S*s learned a greater number of conjunction and preposition pairs than nonsense pairs, 6 *S*s showed no difference, and 4 did better with nonsense material.

The results of Experiment II provide evidence for the facilitating role of meaning per se; they therefore contradict Sheffield's finding that material which is both meaningful and familiar is no easier to learn than merely familiar material. This contradiction is perhaps attributable to Sheffield's conception of meaning which obscures the distinction between a familiar item and one which is meaningful as well.

The significance of the superiority of conjunction and preposition pairs over nonsense pairs will be discussed below (Experiment IV).

Experiment III

Before we can conclude that the results of Experiment II demonstrate the role of meaning (and familiarity) in the *formation* of associations, another possibility must be considered.

Suppose a *S* has associated A with B so that when A is presented he is able to recall B. Köhler (1940) has pointed out that, on the functional level, the reproduction of B must involve two steps. The A item must first be recognized, i.e., the memory trace of A is activated because of its similarity to the present percept of A. The second step is the arousal of trace B (recall by association). Suppose now that the *S* cannot recall B. If there is no difficulty in the recognition of A, how is this failure to produce B to be explained? One possibility is that no bond was formed between A and B in the original learning situation. Another is that the bond was made but for one reason or another the trace of B is not available. For example, if B is a difficult, unfamiliar nonsense item the trace of B may not be readily accessible even though the bond between A and B is well established. The existence of the association could, however, be revealed by the use of a matching test; here, the arousal of B is facilitated by a recognition process. Hence, failure to recall B is not evidence that the association has not been established.

Thus, the superiority of the nouns over conjunctions and prepositions (and of the latter over nonsense syllables) might be explained by reference to the problem of trace availability. Perhaps all associations were equally well formed in the learning process; however, if traces of nouns (in the B position) are more accessible than those of conjunctions, recall scores for the nouns will be higher.[3] If this were the case, then the differences obtained in Experiment II should disappear when learning is tested by the matching procedure. This would be expected since the accessibility of the associated trace is not critical to performance on a matching test. If, however, differences comparable to those obtained with a reproduction test are present with matching also, then we may assume an effect on the formation of associations.

Materials and Procedure

Experiment III was a repetition of the previous one except that *S*s were tested by a matching procedure rather than by recall. The same material was used and the learning conditions were identical. In the matching test, the A item was presented and the *S* was instructed to select the correct B item from a list containing all of the B items arranged in a circular pattern.

Subjects. Thirty-six college students participated, 18 *S*s learned Series I and the other 18 learned Series II.

Results

The results for Experiment III are presented in Table 5 and other details are given in Table 4. The mean number of pairs learned was 4.3, 3.3, and 2.4 for the nouns, conjunctions and prepositions, and nonsense syllables, in that order (Series I and II combined). The standard deviations were 1.51, 1.64, and 1.54, in the same order. The following differences between the means were significant. Nouns vs. conjunctions and prepositions, $t = 4.1$, at the .01 level; conjunctions and prepositions vs. nonsense syllables, $t = 3.0$, at the .01 level.

We may note, then, that although the absolute scores for all materials were significantly higher than in Experiment II, the differences between the materials were maintained on an almost identical level. This near identity of the differences between the means of the various materials for the two experiments suggests that it is not necessary to assume that there is any additional facilitation of meaningfulness and familiarity in the recall process. The influence of these factors seems to reside entirely in the process of forming associations between the two members of a pair.

Experiment IV

In this experiment we wanted to see whether the superiority of the conjunction-preposition pairs over the nonsense pairs (Experiments II and III) was due entirely to their familiarity rather than to the limited meaningfulness they might possess. For this purpose, we compared them with familiarized nonsense syllables.

Method and Materials

The materials consisted of two lists of 6 pairs each. One list was made up of 3 pairs of conjunctions and 3 pairs of prepositions chosen from the lists in Series I and II of Experiment II. The other list consisted of 6 pairs of nonsense syllables also chosen from the lists in Series I and II of

TABLE 5

NUMBER OF PAIRS LEARNED IN EXPERIMENT III

	Series			Comb. Mean	Comb. *SD*
	I	II	I & II		
Nouns	73	82	155	4.3	1.51
Conj. & Prep.	56	62	118	3.3	1.64
Non Syl.	40	45	85	2.4	1.54
N	18	18	36		

Note.—The difference between the noun pairs and the conjunction-preposition pairs was significant at the .01 level ($t = 4.1$). The difference between the nonsense syllable pairs and the conjunction-preposition pairs was also significant at the .01 level ($t = 3.0$).

---

[3] Sheffield's (1946) finding that recall is better when a meaningful item is in the response (i.e., B) position than when it is in the stimulus (i.e., A) position may be explained in this way.

TABLE 6

PAIRS USED IN EXPERIMENT IV

Nonsense Syllables	Conjunctions & Prepositions
daw—lav	off—out
yon—cav	at—from
jor—mot	down—of
vax—fer	how—nor
fij—gos	with—up
ril—gan	now—and

Experiment II. Table 6 contains the materials used in this experiment.

Subjects. Thirty-six college students participated as a group in this experiment.

Procedure. The prefamiliarization procedure was essentially the same as that described in Experiment I with some slight modifications to adjust for a group design. There were two familiarization sessions separated by 48 hrs. At the end of the first session, the *S*s were led to believe that their participation in the experiment was completed. Most expressed surprise when the *E* appeared the second day. After the second prefamiliarization session, one half of the group was asked to leave the room. The remaining group then received one learning trial with the prefamiliarized nonsense syllable now presented in 6 pairs, followed by an interpolated interval of 1 min. filled with work on matchstick problems. *S*s were then tested by being required to reproduce in writing the B item upon being presented with the A member of the pair. When this was completed, the conjunction and preposition pairs were learned and tested for in the same way. Finally, the second group reentered the room and followed the same procedure except that this group learned the conjunctions and prepositions first.

Results

The results for this experiment appear in Table 7. When the data for both groups are combined, the mean number of pairs retained for the two different materials is identical (2.6). The standard deviations were 1.14 and 1.51 for the prefamiliarized syllables and conjunctions-prepositions, respectively. A further look at the individual scores reveals that 17 of the 36 participating *S*s learned a greater number of syllables than conjunctions and prepositions, 15 learned a greater number of conjunction and preposition pairs, and 4 of the 36 *S*s, an equal number of both materials.

Since we have found no difference between familiarized syllables and conjunctions we may assume that the superiority of the latter material in Experiments II and III is attributable solely to its familiarity. The further advantage shown by the nouns in those experiments, therefore, confirms the hypothesis that meaning acts as a separate factor in learning.

Experiment V

The experiments reported so far support the view that the advantage of meaningfulness in learning cannot be attributed to familiarity. The nouns and conjunctions we used occur with approximately equal frequency in the language. But there is another kind of familiarity to be considered—that of certain sequences of words. Although the particular pairs of nouns employed in our experiments may not have been frequently experienced together, it might be argued that they have occurred more often than the pairs of conjunctions. In everyday speech and writing we often list series of nouns so that there is some degree of probability that we have already experienced one or more of the noun pairs used in Experiment II. It is much less likely that any of the conjunction or preposition pairs have occurred previously. Consequently, on the basis of the Miller and

TABLE 7

NUMBER OF CONJUNCTION, PREPOSITION, AND PREFAMILIARIZED SYLLABLE-PAIRS LEARNED IN EXPERIMENT IV

	No. Retained	No. Retained		
Material	Order Syl.-Conj. & Prep.	Order Conj. & Prep.-Syl.	Comb. Mean	Comb. *SD*
Conj. & Prep.	50	42	2.6	1.51
Prefam. Syl.	45	48	2.6	1.14
N	18	18		

Note.—No difference between the means.

Selfridge results, one could predict the superiority of the noun pairs.

(Another possible explanation of the results of Experiment II is provided by the grammatical habit hypothesis.[4] This view stresses conformity with grammatical use rather than the statistical frequency of specific word sequences.)

We do not wish to deny the role of the transitional probability factor in learning; we wish to suggest, however, that it is not sufficient to account for our results. In order to test this view, we need to compare our noun pairs with pairs of words which occur together much more frequently but which cannot be readily organized into conceptual units. For example, the words "here—is" have a high sequential probability but neither word has the meaningfulness of a noun. If the noun pairs were to do better than material of this kind, then the transitional probability approach could not explain the results of our previous experiments.

We decided also to compare pairs such as "here—is" with our conjunction-preposition pairs. The effectiveness of sequential probability would be demonstrated if the new material were learned better.

Since there has been no systematic computation of the transitional probability of word sequences we chose material in the following way: we scanned several books and selected pairs of contiguous words which occurred together quite frequently.

Material and Procedure

Two different series of word pairs were used. In each series there were three lists of six pairs each. List A in each series was composed of material selected as described above. We are confident that the reader examining the pairs of List A will agree that they possess high transitional probability. (See Table 8.) Nevertheless, these words cannot be readily organized into conceptual units. List B consisted of the conjunction-preposition pairs and List C the noun pairs which were used in Experiment II. The experimental procedure was identical with that used in Experiment II. The order in which the three lists were presented was systematically randomized. The order within each list was also randomized from the learning trial to the test and between *S*s.

Subjects. There were 36 college students in the experiment. One half of these learned Series I and the other half Series II.

Results

The results are shown in Tables 8 and 9. For Series I the mean number of pairs learned was 2.3, 2.1, and 3.4 for Lists A, B, and C, respectively. The standard deviations were .79, .89, and 1.12. The difference between the mean obtained with List C and the

TABLE 8

ITEM ANALYSIS OF RESULTS BY EXPERIMENT V

The Number of *S*s Who Responded Correctly to Each Pair

Series I		Series II	
Pair	*N*	Pair	*N*
List A		List A	
are—not	9	before—any	9
such—an	4	here—is	11
will—have	9	and—does	5
am—more	5	was—under	5
which—makes	5	then—it	11
that—this	9	follows—the	11
List B		List B	
down—of	7	now—and	6
then—yet	5	still—if	9
off—out	4	how—nor	4
at—from	6	with—up	4
also—but	9	to—for	5
hence—or	7	in—by	8
List C		List C	
knee—mask	10	match—rent	9
air—chin	10	mat—card	8
fish—rug	10	suit—key	13
sink—park	9	slip—dog	14
wood—shirt	10	boat—wall	15
car—leaf	12	pen—hand	15

---

[4] We wish to thank Susan M. Ervin for calling our attention to this possibility. Personal communication. 1958.

TABLE 9

MEAN NUMBER OF PAIRS LEARNED IN EXPERIMENT V

	Series I		Series II		Combined	
	Mean	*SD*	Mean	*SD*	Mean	*SD*
List A	2.3	.79	2.9	1.09	2.6	.96
List B	2.1	.89	2.0	.91	2.05	.88
List C	3.4	1.12	4.1	.95	3.75	.83

Note.—All of the relevant differences are significant at the .01 level with the exception of $\bar{x}_a - \bar{x}_b$, Series I, which is not significant.

mean obtained for List A and List B is significant at the .01 level. The difference between the means for List A and B is not significant.

For Series II the mean number of pairs learned was 2.9, 2.0, and 4.1 for Lists A, B, and C, respectively. The standard deviations were 1.09, .91, and .95. The mean for List C was significantly better than the mean for List A, and this in turn was significantly better than the mean for List B.

The combined means for both series were 2.6, 2.05, and 3.75 for Lists A, B, and C. The standard deviations for the combined means were .96, .88, and .83. All of the differences were significant at better than the .01 level.

The results give evidence for the role of sequential probability—as shown by the superiority of List A over List B (combined means of Series I and II). But the even greater superiority of the nouns clearly supports the view that the role of meaning cannot be reduced to familiarity of the transitional probability kind.

Experiment VI

We had obtained evidence that meaning plays a role in associative learning above and beyond the facilitation due to familiarity and transitional probability. We then attempted to develop further data in support of Köhler's organization hypothesis.

We chose materials which varied in the extent to which they permitted the formation of conceptual units. Concrete nouns—referring to a specific object or thing, such as "suit," "chair," "dog," etc.—should be easy to build into a unified concept. Abstract nouns—referring to emotions, ideas, etc., such as "grief," "time," "war," etc.—on the other hand, should be more difficult to organize. One can, for example, unify the pair "cat—leaf" by thinking of a cat playing with or eating a leaf. It would be more difficult to do something similar with the abstract word pair "law—time." Our hypothesis predicts, therefore, that concrete noun pairs should be learned more easily than abstract ones. Drawings of specific objects were also employed. We felt that visual representations would be unified easily since the *S* would actually perceive the two objects together. We expected the drawings to do even better than the concrete nouns. We selected verbs as the final class of material. In this case our expectation was that verb pairs would probably not be as easy to organize as concrete nouns. Our hypotheses did not permit a prediction as to the relative difficulty of learning verbs and abstract nouns.

Materials and Procedure

Two series of four lists were constructed. Each series was composed of one list of six pairs of pictures of objects, one list of concrete nouns, one of abstract nouns, and one of verbs. The material is presented in Table 10 and Figures 1 and 2.

Each *S* learned all four lists. Each list was presented for one trial at the rate of one pair every 5 seconds. The *S*s read the pairs or gave the name of the depicted objects aloud under instructions to learn the list in preparation for a test

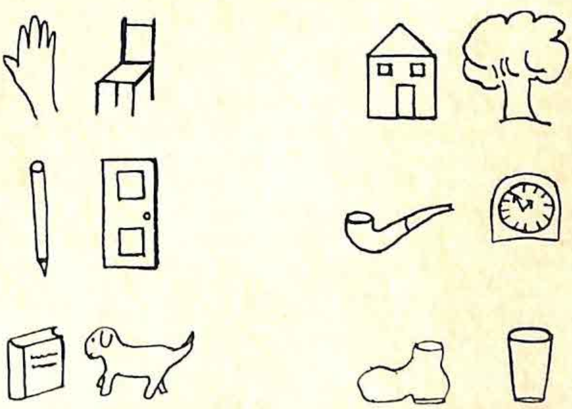


FIG. 1. Pairs of pictures of objects used in Experiment VI, Series A.

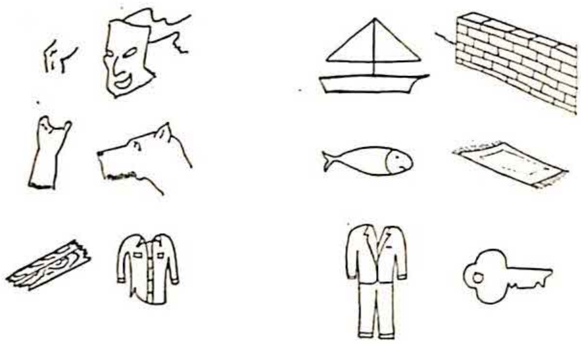


FIG. 2. Pairs of pictures used in Experiment VI, Series B.

in which they would be asked to supply the missing right-hand item upon presentation of the left-hand item. When the learning trial was completed, *E* took 30 seconds to randomize the order of the pairs and to remind *S* very briefly of his task. The test was then given. This procedure was followed for all four lists individually. The order in which the lists were presented was systematically prerandomized so that every list followed and preceded every other list an equal number of times. The order of pairs within the list was randomized from *S* to *S*.

The purpose of the two series (A and B) was to control for possible chance bias in the material. The pictures of Series A were presented as concrete nouns in Series B, and the concrete nouns in Series A were presented as pictures in Series B. In addition, Series B contained completely new lists of abstract nouns and verbs.

Subjects. Forty-eight college students served as *S*s. Twenty-four *S*s learned the material in Series A and 24 *S*s learned Series B.

Results

The results are presented in Tables 10 and 11. The mean number learned for Series A was 3.9, 3.1, 2.3, and 1.9 for the picture, concrete noun, verb, and abstract noun lists, in that order. The standard deviations were 1.34, 1.70, 1.67, and 1.30, in the same order for the four materials.

The mean number learned for Series B was 3.8, 3.0, 2.3, and 2.2 for the picture, concrete noun, verb, and abstract noun lists, in that order. The standard deviations were 1.18, 1.33, 1.52, and 1.34, in the same order.

A comparison of the data of Series A with that of Series B as it relates to pictures and concrete nouns will show that the effects obtained are not produced by any chance bias of material.

When the data for Series A and B are combined, the means were 3.9, 3.0, 2.3, and 2.0 for the pictures, concrete nouns, verbs, and abstract nouns, in that order ($N = 48$). All of the differences were significant except for the obtained difference between the abstract nouns and the verbs. See Table 8 for analysis of variance and *q* test. The results of this experiment confirm our expectations. Performance was enhanced as a direct function of the possibility of forming conceptual units.

TABLE 10

ITEM ANALYSIS OF RESULTS FOR EXPERIMENT VI

The Number of *S*s Who Responded Correctly to Each Pair

Series A		Series B	
Concrete Nouns		Pictures	
suit—key	13	suit—key	17
fish—rug	14	fish—rug	15
knee—mask	10	knee—mask	14
wood—shirt	10	wood—shirt	13
slip—dog	11	slip—dog	15
boat—wall	17	boat—wall	18
Pictures		Concrete Nouns	
house—tree	19	house—tree	14
boot—glass	12	boot—glass	10
pencil—door	13	pencil—door	11
hand—chair	16	hand—chair	12
pipe—clock	17	pipe—clock	12
book—dog	17	book—dog	13
Abstract Nouns		Abstract Nouns	
fame—height	9	death—mood	9
law—time	8	mind—greed	8
truth—grief	7	luck—norm	8
peace—fate	5	pride—age	8
vice—life	9	weight—ideal	9
joy—oath	8	glee—worth	9
Verbs		Verbs	
sang—walked	10	bought—hopped	9
wrote—threw	9	planned—shook	9
brings—melts	8	slept—rang	10
found—jump	9	flows—thinks	9
looks—takes	9	pour—join	9
dives—helps	10	opens—grows	10

One might have thought that the *S* would be able to unify the verb pairs by imagining an individual performing two actions at the same time—e.g., singing while walking. If this did occur it apparently was not as successful in establishing connections as was the conceptualization process involving nouns.

Experiment VII

In this experiment we put the organization thesis to a more direct test. If the hypothesis is correct, then material learned under conditions which encourage the formation of a unit should fare better than material presented under conditions which do not in any special way facilitate the occurrence of unit formation.

Experiment VIIa

Method and materials. Two lists of 6 pairs of concrete noun pairs were constructed. In one list these nouns were joined (or separated) by a third word which constituted an intelligible connective. Thus, we formed a list of three word phrases each of which comprised a unit, e.g., *bank* on *roof*, *lamp* in *bottle*. A second list consisted of the same noun pairs but now joined by an irrelevant connective, that is, one which did not serve the purpose of unit formation, e.g., *bank* late *roof*, *lamp* how *bottle*. In each list three different connectives were used, each serving twice. In the first list the words, on, in, near, were used. In the second list the words how, but, late, were used. (See Table 12.)

The instructions were as follows:

> This is an experiment to find out how well people learn certain materials. I am going to show you a series of six 3-word phrases. In each phrase the first and last words have been underlined. I want you to read aloud the *entire* phrase as I present it to you and try to learn the underlined words in each of the phrases. That is, try to learn the first and last words in each phrase. After you have read through all six phrases once, I will show you the first underlined word of each phrase and ask you to tell me from memory what the second underlined word is in that phrase. You will *not* be asked to recall the middle word. I am only interested in how well you can learn the underlined pairs.

Each phrase was typed on a blank white index card and was presented for a 5-second interval. After one trial was completed 3 seconds were allowed to elapse while *E* randomized the order and made certain the *S* understood his task. The *S* was then shown the first word of each phrase and asked to respond with the correct missing underlined word.

Subjects. Forty *S*s served in this experiment. Group 1, consisting of 20 *S*s, learned the first list, and Group 2, also consisting of 20 *S*s, learned the second list.

Results. The results for this experiment are presented in Tables 12 and 13. The mean number retained for Group 1 was 3.8 (*SD* = 1.41). The mean number retained for Group 2 was 2.7 (*SD* = 1.64). This difference in favor of Group 1 is highly

TABLE 11

NUMBER OF PAIRS LEARNED IN EXPERIMENT VI

	Series A			Series B			Series A & B	
	Number Retained	Mean	*SD*	Number Retained	Mean	*SD*	Number Retained	Mean
Pictures	94	3.9	1.34	92	3.8	1.18	186	3.9
Concrete Nouns	75	3.1	1.70	71	3.0	1.33	146	3.0
Abstract Nouns	45	1.9	1.30	52	2.2	1.34	97	2.0
Verbs	55	2.3	1.67	56	2.3	1.52	111	2.3
N	24			24			48	

Note.—A simple analysis of variance yielded an *F* ratio of 16.6 (an *F* of 2.65 required at the .01 level). A *q* test for the difference between the means yielded the following results (*q* = 0.80 at the .01 level). The difference between the pictures and the concrete nouns was significant (*q* = 0.90). The difference between the pictures and the abstract nouns was significant (*q* = 1.9). The difference between the pictures and the verbs was significant (*q* = 1.6). The difference between concrete nouns and abstract nouns was significant (*q* = 1.0). There was no significant difference between the abstract nouns and the verbs (*q* = 0.30). The difference between the concrete nouns and verbs was significant at the .05 level (*q* = 0.70).

TABLE 12

ITEM ANALYSIS OF RESULTS FOR EXPERIMENT VIIa

The Number of *S*s Who Responded Correctly to Each Pair

Group 1 ($N = 20$)		Group 2 ($N = 20$)	
Material	No. of Corr. Resp.	Material	No. of Corr. Resp.
cake near *road*	13	*cake* but *road*	9
box near *cloud*	11	*box* late *cloud*	8
shoe on *lake*	12	*shoe* but *lake*	8
bank on *roof*	13	*bank* late *roof*	9
lamp in *bottle*	14	*lamp* how *bottle*	9
card in *sand*	13	*card* how *sand*	10
Group 3 ($N = 20$)		**Group 4 ($N = 20$)**	
Material	No. of Corr. Resp.	Material	No. of Corr. Resp.
cake—road	8	*cake* and *road*	7
box—cloud	8	*box* and *cloud*	6
shoe—lake	8	*shoe* without *lake*	9
bank—roof	9	*bank* or *roof*	7
lamp—bottle	8	*lamp* or *bottle*	9
card—sand	7	*card* without *sand*	8

significant ($t = 2.44$, significant at the .01 level).

Before we accept these findings as evidence for the hypothesis, we must consider a possible objection. It can be argued that the conditions of learning for Group 2 actually inhibited performance and thus do not provide a fair test of our thesis. The middle word for Group 2 may have interfered with the link between the first and last words, thus lowering the scores. We may ask why the middle word did not have this effect for Group 1. An obvious answer would be that for this group the middle word helped the learner form a unit and thus enhanced instead of inhibited the learning. This, however, is precisely the point.

TABLE 13

NUMBER OF PAIRS LEARNED IN EXPERIMENT VIIa

	N	No. Retained	Mean	*SD*
Group 1	20	76	3.8	1.41
Group 2	20	54	2.7	1.64
Group 3	20	48	2.4	1.32
Group 4	20	46	2.3	1.18

Note.—Group 1 vs. 2: $t = 2.44$; significant at the .05 level. No difference between Group 2 and Group 3 ($t = 0.60$). The difference between Group 1 and Group 4 is significant at the .01 level.

To check on the validity of this criticism, we decided to run a control group which would learn the same pairs as Group 1 and Group 2 without the presence of any kind of connective between the pair members (Group 3). If the irrelevant connective served to interfere with learning or recall, then eliminating the connective ought to make the control group performance superior to the performance of Group 2.

Twenty adults drawn from the same population as Groups 1 and 2 served as the *S*s for this control group. The same learning conditions were maintained except for a change in the wording of the instructions to eliminate sections which did not pertain to this design.

The mean number of pairs retained was 2.4. The standard deviation was 1.32. The difference between the means for the Control group and Group 2 was not significant ($t = .60$).

A second objection might be raised on the basis of grammatical habits. Although the particular three-word sequences do not occur very often, the connectives used in the word pairs learned by Group 1 created a triad more in accordance with grammatical structure than was the case for Group 2. For example, the phrase "lamp in bottle" is a more natural grammatical sequence than "lamp how bottle." Moreover, the triads for Group 1 have a greater transitional probability than those for Group 2; some of the latter would represent a zero-order approximation to English (cf. Miller & Selfridge). Thus, both the grammatical habit and sequential probability hypotheses predict that the Group 1 material would be learned more readily. No reference to meaning would be necessary.

To test this possibility, we repeated the experiment with the same word pairs but employed different words to connect them. The connectives were chosen so that the triads would be as grammatically natural and would have the same transitional probability as the material used for Group 1. However, these connectives would not aid the formation of conceptual units. The sequence "lamp or bottle" is as good grammatically and is as probable as "lamp in bottle." But according to the organization viewpoint, the word "or" does not facilitate unit formation.

Materials and method. The materials used with Group 4 are shown in Table 12. The procedure was identical with that used for Groups 1 and 2.

Subjects. Group 4 consisted of 20 college students.

Results. The mean number of pairs learned was 2.3 with a standard deviation of 1.18. This result should be compared with the mean of 3.8 obtained by Group 1. The difference of means is significant at the .01 level. Furthermore, Group 4 does no better than Groups 2 or 3.

The experiment gives a decisive answer. Whatever role one assigns to grammatical habits and transitional probability in verbal learning, it is clear that these factors cannot fully account for the present results. The possibility of unit formation seems to be the crucial factor.

Experiment VIIb

In this experiment our purpose and design was similar to Experiment VIIa, except that pictures were used. Group 1, consisting of 20 *S*s, learned six pairs of outline pictures of familiar objects, e.g., card—glass. The pictures were drawn so that each pair constituted a unit. Thus, a "card in a glass" was presented. See Figures 3 and 4 for reproductions of the pictures used in this experiment.[5] Group 2, consisting of 20 *S*s, learned the same six

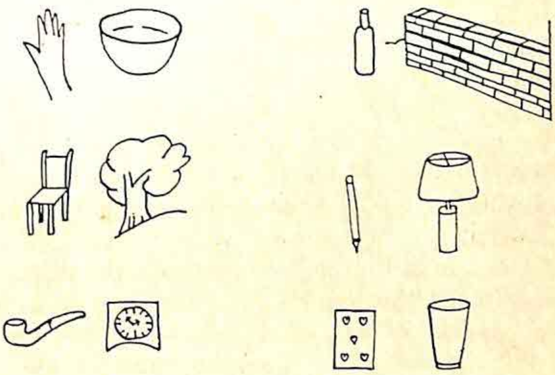


FIG. 3. Pairs learned by Group 2, Experiment VIIb.

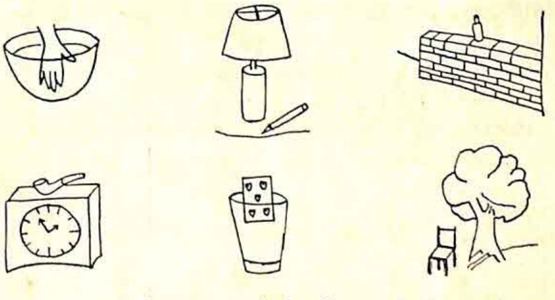


FIG. 4. Pairs learned by Group 1, Experiment VIIb.

---

[5] This technique of joining items into an integral unit is analogous to one first employed by Asch, Ceraso, and Heimer (1960).

pairs of pictures. For this group, however, the members of each pair were presented in the traditional paired-associate manner.

The instructions given to *both* groups were:

> You will be shown six pairs of pictures of things you have seen often. I want you to try and learn these pictures, that is, try and learn which pictures are paired together. After we have gone through the list once, I will show you one of the pair and ask you to tell me verbally what you remember the missing partner to be.

Each pair was presented for 5 seconds, and a 30-second interval followed the last presentation. The test order was randomized.

Results. The results for this experiment are presented in Table 14. Group 1 learned a mean number of 4.9 pairs ($SD = 1.35$). Group 2 learned a mean number of 3.3 pairs ($SD = 1.26$). This difference is highly significant ($t = 3.6$, significant at the .01 level).

DISCUSSION

Our results show that pairs of familiar items are easier to associate than pairs of novel ones. Why this should be so is not entirely clear but one reason is surely the fact that with familiar material the *S* has less to learn. Since he already knows the individual items, he is free to concentrate on forming a bond between them.

In his discussion of the organization hypothesis, Köhler does not mention the role of familiarity. It is clear, however, that emphasis on meaning is perfectly compatible with acceptance of the importance of familiarity as well. The real issue, we believe, is whether the effect of meaningfulness in learning is reducible to familiarity of any kind. In the series of experiments reported above we have attempted to prove that such a reduction is not possible. Meaning appears to be an independent factor.

Although meaning is a result of past learning it enables the *S* to go beyond what he has learned. Similarly, while insight in problem solving may be based on past experience, it is not simply a reproduction of the past but represents something new. When the *S* knows the meanings of words he can "play" with them and put them into new combinations. This possibility makes them easier to associate than nonsense syllables or familiar but less meaningful items. When we refer to these combinations as conceptual units we do not wish to imply that such units are sensible or logical. They are conceivable or imaginable.

TABLE 14

RESULTS OF EXPERIMENT VIIb

Pairs of Pictures Learned

Material	Group 1 ($N = 20$)	Group 2 ($N = 20$)
pipe—clock	18	11
card—glass	15	7
chair—tree	18	11
bottle—wall	16	12
pencil—lamp	15	10
bowl—hand	16	11
Total	98	66
X	4.9	3.3
SD	1.35	1.26
$t_{\bar{x}_1 - \bar{x}_2}$	3.6**	

** Significant at .01 level.

Many investigators, in recent years, have perhaps too hastily concluded that the results of their experiments make unnecessary any reference to the role of meaning as we have described it. But as we have already stated, evidence in favor of a habit explanation does not prove that this is the only possible explanation.

For example, according to Noble (1952a) "meanings are habits" (p. 429). The greater the number of responses conditioned to a particular stimulus the more meaningful is this stimulus. He advances an operational index of stimulus meaning which is the "mean frequency of continued written associations made by subjects within a 60 sec. time interval" (1952a, p. 429).[6] Noble assumes that this index (*m* value) is an important variable in verbal learning. A serial learning experiment (1952b) showed that *m* value was positively correlated with ease of learning. Presumably the facilitative effect of high *m* value would follow from the associationist explanation outlined above on page 1.

If we were to apply Noble's *m* value (or Mandler & Huttenlocher's [1956] *f* value) index to our material, it might be assumed that concrete nouns would elicit more associations than conjunctions or abstract nouns. Informal tests confirm this assumption. Thus, the superiority of concrete nouns might be explained as the outcome of high *m* values. Yet our explanation cannot be ruled out, unless one can find a situation for which the two hypotheses make different predictions. Noble's hypothesis would predict that the concrete nouns should be learned more readily, not only as paired associates but in serial learning as well. On the basis of the organization view, we would argue that, in the learning of a series, the formation of conceptual units would be quite difficult, since the words are presented in a continuous sequence. Therefore, the concrete nouns might do no better than other kinds of material, especially when only one or two presentations are given. Preliminary data obtained by John Catalano[7] yield the following means for the number of items recalled by 23 *S*s after one learning trial each (using a test of free recall) with a list of 12 concrete nouns, a list of 12 conjunctions and prepositions, and one of abstract nouns: 8.9, 8.7, and 8.0, respectively. These results are contrary to the *m* value hypothesis. Further research on this problem is in progress.

To explain the superiority of picture pairs over the other kinds of material and the superiority of concrete nouns over verbs found in Experiment 6, in terms of associative value, it would be necessary to obtain corresponding differences in *m* value. This seems unlikely. Moreover, the fact that the pairs in List A of Experiment 5 are not as well learned as noun pairs argues against the Mandler and Huttenlocher interpretation. In this case, the words have been linked together frequently so that no mediating bridges would be needed to aid in the formation of a bond.

While Noble and Mandler and Huttenlocher agree that meaning, as defined by them, does benefit learning, Sheffield (1946) finds no facilitative effect. This failure may be an outcome of the way in which he attempted to make a term meaningful. Attaching such responses as blacking out the word or putting a staple through it does not provide the kind of meaning that would favor the formation of associations.

Another factor in the effect of meaningfulness which we have not discussed has been suggested by Sheffield. Meaning, as an internal response, facilitates learning because of its role in making for stimulus differentiation. The words "cup" and "cut" as sheer sounds are very similar. But the unique internal responses attached to each differentiates the two items and thus operates to reduce interference. However, Sheffield's own data do not support this hypothesis since meaning did not yield any benefit in addition to familiarity. Nor does this view account for the differences obtained in Experiments 6 and 7.[8]

---

[6] Noble eliminates clang and tangential associations unrelated to the meaning of the stimulus term. This procedure shows that *m* value is not identical with meaning; these associations are dependent on *S*s knowledge of the meaning of the word. It is not clear what associations to a nonsense syllable signify.

[7] Brooklyn College.

Another example where results supporting their hypothesis led the investigators to conclude that meaning is not an independent factor is provided by the Miller-Selfridge experiment. These authors emphasize their finding that some sequences of a third or fourth order of approximation are learned as well as text passages. They interpret this result as evidence that learning depends not on the meaningful nature of the sequence but on the presence of familiar short range associations between words.

It is interesting to find that this experiment was repeated with some modification by Marks and Jack (1951). Instead of testing by means of free recall, they required the *S*s to reproduce the words in the same order in which they had been presented during learning. They argue that this method provides a better test of the Miller and Selfridge hypothesis which stresses short range association. In contradition to the earlier experiment, Marks and Jack found that "recall seems to be very much a function of the *a priori,* relative meaningfulness of the materials" (p. 300).

Intuitively, it did not seem possible to us that text passages would not be learned more easily than some of the lower orders of approximation. Therefore, we selected two passages from the Miller and Selfridge material which according to their result were learned as well as text and repeated their experiment with the identical procedure. The results for 16 college students were:

1. Twenty-word passage, third-order approximation—52% recall.

Twenty-word text passage—78% recall.

2. Thirty-word, fourth-order approximation—58% recall.

Thirty-word text passage—72% recall.

We have no explanation for the discrepancy but our results, together with those of Marks and Jack, suggest that Miller and Selfridge have not ruled out meaning as a factor in verbal learning. It should be pointed out, in addition, that their own data might be explained in terms of "short range meanings" as well as short range associations.

One final example: grammatical sequences of words can be learned better than random sequences of the same length. This may be a result of transfer from previously established grammatical habits. However, a grammatical sequence is usually meaningful (of course in a more complex sense than is the case for an isolated word) and the role of meaning may be independent of grammatical habits. Again, evidence in support of one factor does not eliminate the other. Experiments analogous to the ones reported in this paper could perhaps assess the separate effects of each. For example, *S*s might be given the task of learning two passages equally well structured grammatically but differing in meaningfulness. (Perhaps a passage from Hegel and one from Bertrand Russell?)

It is important to clarify the sense in which we use the term "organization" in connection with Köhler's hypothesis. We would agree with Postman and Riley (1957) that in the case of verbal learning, the term "organization" as it is used in perception, with its implications relating to a specific brain model, is inappropriate. We use it in the present context only to refer descriptively to the possibility which Köhler discusses, namely, that the learner finds it easy to relate meaningful items to one another, to imagine the component items in some relation or juxtaposition to one another. Although this is intuitively understandable, further research is required to clarify why it is easier to organize meaningful items than other types of items.

---

[8] Another possible explanation of some of our findings, which is suggested by Sheffield, is that meaningful material more readily lends itself to that type of mnemonic device involving intermediary links. Thus, for example, "pen—hand" may suggest "pen—write—hand" to the learner. It is doubtful, however, that this could occur or did occur very often. Furthermore, it is not obvious how the results of Experiments V and VI would follow from this view.

SUMMARY

The roles of familiarity and meaning in associative learning were investigated.

Experiment I revealed the effect of familiarity: prefamiliarized nonsense syllables were learned better than novel ones.

In Experiment II a group of *S*s learned three lists of six pairs of words: a list of nouns, a list of three conjunction pairs and three preposition pairs, and a list of nonsense syllable pairs. The nouns were considered to be meaningful and familiar, conjunctions and prepositions primarily familiar, and syllables neither familiar nor meaningful. The results showed that nouns were better learned than conjunctions and prepositions, and these in turn were better learned than the syllables.

Experiment III was an attempt to localize the influence of meaning and familiarity. Experiment II, therefore, was repeated, this time using matching instead of reproduction to test learning. Although the absolute level of performance was enhanced as compared with Experiment I, the relative differences between materials remained the same. We concluded that in our experiments meaning and familiarity operated almost entirely to facilitate the formation of the associations.

Experiment IV was designed to determine whether the advantage of conjunctions and prepositions in Experiments II and III is due to familiarity or to limited meaning. Prefamiliarized nonsense syllables were learned as well as conjunctions and prepositions, showing that familiarity was the crucial factor.

Experiment V provided evidence relating to the role of sequential probability in our experiments. The results indicated that while this factor does facilitate learning to some extent, it is not sufficient to account for our earlier findings.

In Experiment VI we tested the possibility that meaningful material allows for the formation of conceptualizable units which combine the separate items into a new whole. The *S*s learned four lists composed of six pairs each: a list of pictures of real objects, a list of concrete nouns, a list of verbs, and a list of abstract nouns. We assume that these varied in respect to the readiness with which they could be formed into novel units. The results showed that pictures were learned better than concrete nouns which in turn were superior to verbs and abstract nouns. The verbs and abstract nouns did not differ significantly.

In Experiment VIIa we investigated further the hypothesis that the conceptualization of objects is facilitating because it encourages the formation of units. Two groups of *S*s learned six pairs of concrete nouns. Group 1 learned the pairs as parts of a three-word phrase in which the middle word was an intelligible connective under conditions which encouraged unit formation. Group 2 learned the same pairs as parts of a three-word phrase where the middle word did not reasonably connect the two members of the pair. Performance for Group 1 was superior to Group 2. A third group served as a control. This group learned the same six pairs but with no connective whatsoever. The performance of this group was not significantly different from Group 2. A fourth group learned phrases which formed sound grammatical units but did not form good conceptual units. This group was not significantly different from the control.

In Experiment VIIb we used six pairs of pictures of familiar objects. For Group 1 the two pictures of each pair were presented as one unit. For Group 2 the pairs were presented in the traditional paired-associate manner. Group 1 was highly superior to Group 2.

These experiments were considered in their relation to the explanations offered to account for the role of meaningfulness.

REFERENCES

Asch, S. E., Ceraso, J., & Heimer, W. Perceptual conditions of association. *Psychol. Monogr.*, 1960, **74**, (3, Whole No. 490).

Chomsky, N. Syntactic structures. The Hague: Mouton & Co., 1957.

Glaze, J. A. The association value of nonsense syllables. *J. genet. Psychol.*, 1928, **35**, 255–267.

Fries, C. F. *The structure of English*. New York: Harcourt, Brace, 1952.

Hovland, C., & Kurtz, K. H. Experimental studies in rote-learning theory: X prelearning syllable familiarization and the length-difficulty relationship. *J. exp. Psychol.*, 1952, **44**, 31–39.

Köhler, W. *Gestalt psychology*. New York: Liveright, 1929.

Köhler, W. *Dynamics in psychology*. New York: Liveright, 1940.

Mandler, G., & Huttenlocher, J. The relationship between associative frequency, associative ability and paired associate learning. *Amer. J. Psychol.*, 1956, **69**, 424–428.

Marks, M. R., & Jack, O. Verbal context and memory span for meaningful material. *Amer. J. Psychol.*, 1952, **65**, 298–300.

McGeoch, J. A., & Irion, A. L. *The psychology of human learning*. New York: Longmans, 1952.

Miller, G. A., & Selfridge, J. A. Verbal context and the recall of meaningful material. *Amer. J. Psychol.*, 1950, **63**, 176–185.

Noble, C. E. An analysis of meaning. *Psychol. Rev.*, 1952, **59**, 421–430. (a)

Noble, C. E. The role of stimulus meaning (m) in serial verbal learning. *J. exp. Psychol.*, 1952, **43**, 437–446. (b)

Osgood, C. E. *Method and theory in experimental psychology*. New York: Oxford Univer. Press, 1953.

Osgood, C. E. A behavioristic analysis of perception and language as cognitive phenomena. In *Contemporary approaches to cognition: A report of a symposium at the University of Colorado, May 12–14, 1955*. Cambridge, Mass.: Harvard Univer. Press, 1957.

Postman, L., & Riley, D. A. A critique of Köhler's association theory. *Psychol. Rev.*, 1957, **64**, 61–72.

Robinson, E. S. *Association theory today*. New York: Appleton-Century, 1932.

Sheffield, F. D. The role of meaningfulness of stimulus and response in verbal learning. Unpublished doctoral dissertation, Yale Univer., 1946.

Thorndike, E. L., & Lorge, I. *The teacher's word book of 30,000 words*. New York: Bureau of Publications, Teachers College, 1944.

Waters, R. H. The law of acquaintance. *J. exp. Psychol.*, 1939, **24**, 180–191.

(Received March 20, 1958)

Vol. 74, No. 5 Whole No. 492, 1960

Psychological Monographs: General and Applied

RELATIVE EFFECTIVENESS OF PSYCHOTHERAPEUTIC PROGRAMS:

A MULTICRITERIA COMPARISON OF FOUR PROGRAMS FOR THREE DIFFERENT PATIENT GROUPS[1]

G. W. FAIRWEATHER,[2] R. SIMON,[3] M. E. GEBHARD, E. WEINGARTEN,
J. L. HOLLAND,[4] R. SANDERS,[5] G. B. STONE, and J. E. REAHL

Veterans Administration Hospital, Perry Point, Maryland

PSYCHOTHERAPISTS in neuropsychiatric hospitals are continually faced with the problem of formulating feasible and economical psychotherapeutic programs for a heterogeneous population of patients. Historically, such programs have usually been established upon the theoretical propositions accepted by those in positions of authority and, frequently, little attempt has been made to evaluate their effectiveness. In addition, when such evaluations have been carried out, a single criterion is selected and its relationship to other critieria has been ignored. The present study attempts to compare the relative efficiency of four common psychotherapeutic programs for three patient groups using a number of diverse inventories, projective devices, and rating scales.

Probably the most common treatment program in neuropsychiatric hospitals today provides each patient with an individual work assignment and planning for post-hospital adjustment. This program is based upon the assumption that the hospital work situation can be therapeutic. For that reason, the hospital personnel attempt to place the patient in a work situation suited to his unique skills and needs. This opportunity for maximal use of his abilities in a job setting may enhance the patient's self-esteem and result in more adaptive behavior in the extramural setting.

A second common treatment program employs individual work assignments, individual psychotherapy, and departure planning. This assumes that underlying conflicts must be resolved before any effective personality change can occur and, further, that personality change is essential for extramural adjustment. It assumes too that the relationship between the psychotherapist and patient is the most effective instrument for resolving unconscious conflicts. A host of publications and documents signify that individual psychotherapy has, indeed, become one of the most common psychotherapeutic approaches of our time.

A third treatment program uses individual work assignments, group psychotherapy, and departure planning. It differs from the two preceding programs only with respect to group psychotherapy. Proponents of group psychotherapy, as the medium for personality change, maintain that, since man has learned his maladaptive responses in a social milieu, the most effective vehicle for changing such responses is group psychotherapy. In this program the situation for personality change is the pattern of interpersonal relationships occurring in group therapy.

---

[1] The authors wish to express their appreciation to the entire staff of VA hospital, Perry Point, Maryland, for their assistance in making this study possible. We are particularly indebted to Sarah Shifflet, Anna Wirsing, Grace Pruitt, and James Sprouse, nursing assistants, and Mary McMenamin, nurse, who served on the experimental ward. For clerical and statistical help we are grateful to Ralph Forsyth, Barbara Creswell, Kathryn Boyd, Itsuye Sakai, L. Arlene Zealor, Marian Barry, and Lucy Yoshizawa. To Helen Jefferys, Marge Edelson, and John Webster we express gratitude for their social work services. We are especially indebted to Earl Brannon, Manager, and Willy Oppler, Chief of Professional Services, for their administrative assistance in carrying out this project. We are grateful to Roy E. Buehler for his contributions to the theoretical aspect and Laura Kent for her aid in preparation of the manuscript.

[2] Now with VA Hospital, Palo Alto, California.

[3] Now with National Institutes of Health, Bethesda, Maryland.

[4] Now with National Merit Scholarship Corporation, Evanston, Illinois.

[5] Now with Philadelphia State Hospital, Philadelphia, Pennsylvania.

The fourth program, albeit used less frequently than the preceding three, is a social interaction, or "group living therapy." This rests upon the proposition that the total life situation of the patient can be used as an instrumentality for personality change. Therefore, the patients live and work together, and, as an integral part of this program, participate in group psychotherapy. This program differs from the other three in that all its functions are group oriented. The rationale is that changes in maladaptive perception and subsequent behaviors are best effected by planned social situations which provide an opportunity for correcting inappropriate interpersonal responses. Each situation provides an atmosphere for group decision and responsibility conducive to group cohesiveness. It is the cohesiveness of the group which mirrors the accepted norms of the society and provides the background for corrective emotional experiences. In this atmosphere, all group activities are believed to be potentially therapeutic.

In general, psychotherapeutic programs such as those described have been more successful with neurotic and acute psychotic patients than with chronic psychotic patients.

From the aforementioned experience and rationale two general hypotheses evolve. They are:

1. Treatment programs involving psychotherapy effect more change in the direction of adaptability than work programs alone.
2. Nonpsychotics and short-term psychotics change more in the direction of greater adaptability with any treatment program than long-term psychotics.

PROCEDURE

The experiment took place on a ward in a large Veterans Administration neuropsychiatric hospital. The number of patients on the ward at a time varied from 40 to 62. The ward staff included a psychiatrist, two psychologists, a social worker, a psychological intern, a nurse, and four nursing assistants. As for its physical features, there were three large dormitories, a recreation room, a reading and visitors room, and a television room. All patients in the study remained on this ward for the duration of their treatment.

The general ward program for all patients in the experiment included responsibility for ward housekeeping and, in addition, two recreational evenings per week. The first of these took place on Sunday or Monday and included a structured recreational situation. During these periods a member of the special services division of the hospital, along with a group of women volunteers from the community, planned recreational situations in advance, i.e., dances, table games, etc. The second recreational situation, on Thursday evenings, was an unstructured situation; the women volunteers arrived on the ward and, together with the patients, planned recreational activities for the evening. Usually informal groups were formed spontaneously. They played cards, danced, and engaged in similar activities. In addition, the ward recreation rooms were open from 11 to 12 in the morning and from 3 to 9 o'clock in the evening. When the patients were not participating in their specific treatment programs, they could watch TV, play pool or cards, read, and engage in other activities. During this time a nursing assistant was on duty.

The Four Experimental Conditions

In the first treatment program, a rehabilitation team[6] provided each patient with an individual work assignment and planning for posthospital living. This group constitutes the control group for the study and will be referred to as Group C. For the second group (Group I), the treatment program consisted of an individual work assignment, individual psychotherapy, and departure planning outlined by the rehabilitation team. The treatment program for the third group (Group G) included an individual work assignment, group psychotherapy, and departure planning by the rehabilitation team. For the fourth group (Group GG), the treatment consisted of a group work situation, group living situation, and group psychotherapy.

---

[6] Consisting of a psychologist, psychiatrist, counseling psychologist, social worker, and psychological intern. Any further use of the term "rehabilitation team" refers to these individuals.

Upon their arrival on the ward, members of the C, G, and I groups were interviewed by a counseling psychologist and given individual work assignments. In addition, the G and I groups were given therapy appointments. Groups C, G, and I were responsible for ward housekeeping in all dormitories except the one occupied by the GG group. For these three groups, each patient was assigned to ward housekeeping for a period of two weeks. This was done on a rotating basis.

The GG group lived together in a single ward dormitory. Their work assignment was to produce a hospital newspaper every two weeks. Each afternoon at 1 o'clock they reported to a room which had been set aside as the newspaper workroom. Here, under the supervision of a staff member, they had the total responsibility for running the newspaper: they made decisions about the content of the paper, the reporting of the articles, its reproduction, and all other matters relating to the administration and design of the paper. The first hour of each afternoon was devoted to a discussion of their work. In addition, they received group therapy and were responsible as a group for the housekeeping in their dormitory.

Therapists and Therapy

Two staff psychologists served as therapists for the patients in the G and GG groups. One of these had four years of postdoctoral psychotherapeutic experience and the other had two years. The 48 patients in these two groups were divided evenly between the therapists.

The individual therapy patients (Group I) were carried by 15 psychotherapists with a wide range of experience. Five clinical psychology staff members, whose range of experience varied from two to five postdoctoral years, provided individual psychotherapy. One staff psychiatrist, with six years of psychotherapeutic experience, carried an individual therapy patient. In addition, four clinical psychology trainees and four psychiatric residents provided individual psychotherapy. All were supervised by consultants.

Both the group psychotherapy and individual psychotherapy sessions were psychoanalytically oriented. Group sessions were held for 1—1/2 hours two times weekly and individual therapy sessions varied from two to four per week depending upon the judgment of the psychotherapist. Groups averaged six patients and ranged from four to eight. The groups were open groups and usually consisted of neurotics, acute psychotics, and chronic psychotics.

Criterion for Discharge or Failure

At the weekly meeting of the rehabilitation team, each patient's current hospital adjustment was discussed and his readiness for discharge was evaluated. No patient left the treatment program unless it had been unanimously agreed that he was ready for discharge. In addition to the 85 patients discharged, there were 11 who were considered failures. The criterion for failure was the team's decision that the patient had shown no progress, or had shown a decrement in his general adjustment after six months of treatment. Such patients were considered failures and were sent from the experimental ward to other hospital wards where a more appropriate treatment program was possible.

Sampling Procedures

When the program started, each patient coming to the ward was randomly assigned to one of the four treatment groups. After a sufficient sample in each group had been accumulated by random assignment, new patients were placed by matching across treatments on age, diagnosis, and length of hospitalization. When new patients who did not meet the matching criteria arrived on the ward, they were randomly assigned to one of the four treatment groups. Thus, over a period of time a treatments by diagnoses factorial analysis of variance design was attained with 24 patients in each of the four treatment groups and 32 patients in each of the three diagnostic groups.

The four treatments have been discussed. The three diagnostic groups were: (*a*) nonpsychotic, (*b*) short-term psychotic, and (*c*) long-term psychotic.[7] Each patient in the nonpsychotic group had been diagnosed a nonpsychotic. Each patient in the short-term psychotic group had less than one year of previous hospitalization in addition to his psychotic diagnosis. Each patient in the long-term psychotic group had a psychotic diagnosis and previous hospitalization exceeding one year.

A descriptive comparison of the treatment and diagnostic groups appears in Table 1 and 2, respectively.

Table 1 indicates that the treatment groups are quite homogeneous with respect to diagnosis, amount of drugs received, age, education, race, length of previous hospitalization, pensions for NP illness, age of first hospitalization, employment history, occupational status, and marital status.

On the other hand, Table 2 reveals significant difference among the three diagnostic groups. The stratifying variable of previous weeks of NP hospitalization significantly differentiates the three diagnostic groups as do the pension, employment, and marital status variables. No significant differences are found for age, education, race, age of first NP hospitalization, weekly salary, occupational status, or percentage of patients on tranquilizing medications. Viewed collectively, the differences suggest that the long-term psychotic is more maladaptive than either the nonpsychotics or short-term psychotics.

Experimental Instruments and Testing Situation

In selecting criteria for evaluating change, the experimenters attempted to include measures that were frequently used for this purpose and encompassed a broad range of behavior. The areas selected for assessment included individual behavior, self-evaluation, fantasy, and follow-up. Specifically the instruments were: Ward Behavior, MMPI, Q sort, Holland Vocational Preference Inventory (HVPI), TAT and a Follow-up Rating

TABLE 1

COMPARISON OF THE FOUR TREATMENT GROUPS ON DEMOGRAPHIC CHARACTERISTICS

Variable	GG	G	I	C	Test of Significance
Matching Variables					
Diagnoses					
% character disorders	12.5	12.5	12.5	12.5	
% neurotics	20.8	20.8	20.8	20.8	$\chi^2 = 0$
% psychotics	66.7	66.7	66.7	66.7	
Mean age	36.25	33.17	34.21	34.92	$F = \ldots$[a]
Mean weeks of prior NP hospitalization	51.6	49.2	53.8	54.8	$F = \ldots$
Concurrent Variable					
% receiving tranquilizing drugs	54	58	58	42	$\chi^2 = 1.58$
Demographic Variables					
Education					
Mean grade completed	10.5	10.6	11.0	11.2	$F = \ldots$
Race					
% white	79	75	75	75	
% Negro	21	25	25	25	$\chi^2 = .17$
% pensioned for NP illness	67	50	67	58	$\chi^2 = 1.52$
Mean age of first NP hospitalization	28.5	27	29	31	$F = \ldots$
Employment					
Mean % time employed in past 10 years	71.5	75	68	71	$F = \ldots$
Mean weekly salary of last job	74.3	63.2	68.5	62.7	$F = \ldots$
Occupational status					
% professional-managerial	12.5	20.8	25	16.7	
% skilled trades, sales, clerical	33.3	20.8	20.8	20.8	$\chi^2 = 2.48$
% unskilled workers	54.2	58.4	54.2	62.5	
Marital Status					
% currently married	25	33	25	25	
% single or divorced	75	67	75	75	$\chi^2 = .53$

[a] The F ratio value . . . indicates that the within-group variance was greater than the between-group variance.

---

[7] Diagnostic groups are listed in the tables as follows: nonpsychotics (N), short-term psychotics (Ps), and long-term psychotics (Pl).

TABLE 2

Comparison of the Three Diagnostic Groups on Demographic Characteristics

Variable	N	PS	PL	Test of Significance
Matching Variables				
Mean age	36.09	34.41	33.31	$F = \ldots$[a]
Mean weeks of prior NP hospitalization	25.34	20.47	111.22	$F = 45.91$**
Concurrent Variables				
% receiving tranquilizing drugs	37.5	56.3	65.6	$\chi^2 = 5.27$
Demographic Variables				
Education				
Mean grade completed	11.03	10.34	11.09	$F = \ldots$
Race				
% white	81.3	65.6	81.3	$\chi^2 = 3.94$
% Negro	18.7	34.4	18.7	
% pensioned for NP illness	37.5	59.4	90.6	$\chi^2 = 19.46$**
Mean age of first NP hospitalization	31.19	29.75	25.88	$F = 3.06$
Employment				
Mean % time employed in past 10 years	79.22	83.50	51.09	$F = 17.02$**
Mean weekly salary of last job	68.06	73.72	56.13	$F = 2.37$
Occupational status				
% professional-managerial	12.5	21.9	21.9	$\chi^2 = 3.32$
% skilled trades, sales, clerical	34.4	18.8	18.8	
% unskilled workers	53.1	59.3	59.3	
Marital status				
% currently married	28.1	46.9	6.2	$\chi^2 = 13.39$*
% single or divorced	71.9	53.1	93.8	

[a] The F ratio value . . . indicates that the within-group variance was greater than the between-group variance.

*.01 level of significance.

**.001 level of significance.

Scale.[8] Their construction and scoring will be discussed in subsequent sections devoted to the findings obtained with each assessment device.

Every patient in the study was tested twice with the instruments listed above. The first assessment was made at the time the patient transferred to the experimental ward and the second shortly before he left the ward. The tests were administered during scheduled sessions in an office building which was located some distance from the ward. The same examiner administered both initial and terminal tests. The initial testing occurred, with a few exceptions, on the first Monday afternoon and Tuesday morning following a patient's transfer to the experimental ward. At that time the patient was instructed by the ward nurse to report to the office building "for testing." In a few cases, due to uncontrollable circumstances such as illness, the time interval between transfer to the ward and testing was somewhat longer, though it never exceeded two weeks. The second administration usually was scheduled for the Monday afternoon and Tuesday morning of the week during which the patient was to leave the experimental program. Here again there were occasional exceptions.

The tests were administered to groups which averaged three or four patients and ranged from two to nine. Groups undergoing initial and terminal testing worked in different rooms.

The following instructions were given for the initial testing:

> All of you have recently been transferred to Ward 19A from other parts of the hospital. Part of the routine procedure connected with transfer to 19A is some psychological testing and this is why you were asked to come over here today. The purpose of these tests is to provide information about you which will be of help to your doctors on the ward in helping you. It is therefore in your interest to cooperate to the best of your ability in doing these things which you will be asked to do, and I hope you

---

[8] The Ward Behavior Scale, Q sort, and Follow-up Scale have been deposited with the American Documentation Institute. Order Document No. 6239 from the A.D.I. Auxiliary Publication Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance $2.50 for photoprints or $1.75 for 35 mm. microfilm. Make checks payable to Chief, Photoduplication Service, Library of Congress.

> will do so. Now, some of you may have done some of these things before, either when you first came to Perry Point, or at some other time. But, even so, even if you have done them only a few days ago we would like you to do them again, since it is important that we get the most accurate picture that we can of you as you are now, as you are entering 19A. There are several things which you will be asked to do and these will take, probably, all of the afternoon and part of tomorrow morning. Unless there are any questions, we will begin.

The examiner then distributed the Q sort cards and place cards, drawing the most-like-6 and least-like-6 cards out of the stack and placing them separately in front of the patient. He then proceeded with the instructions as follows:

> I have just placed before each of you a stack of cards, each of which has a typewritten statement on it. What I would like you to do is to sort these cards according to the way the statements apply or don't apply to you. The way I want you to start is to go through the whole stack, reading all the statements. Out of these cards you are to select the six which are most like you the way you are, which are most true of you, and to put them on the card which says most-like-6. Also select the six statements which are least like you the way you are, which are least true of you, and put them on the card marked least-like-6. Now remember, in doing this you are supposed to go through the whole stack of cards.

After a patient had completed the first sort, the examiner counted the cards he had placed in each pile, attached them and placed them in a separate envelope for each patient. Then he drew the most-like-9 and least-like-9 cards from the stack and said:

> Now go through the remainder of the cards and, out of all the cards you have left, select the nine statements which are most like you and the nine which are least like you and put them on these cards. Again, go through all the cards you have left.

This sorting was repeated twice more with 15 and 20 cards each.

After a patient had completed the Q sort and the examiner had checked the numbers in each pile, an HVPI booklet and answer sheet were placed before him and the examiner said:

> The next thing you will do is to go through this list of jobs or occupations and indicate on the answer sheet your first impression as to whether a particular job or occupation appeals to you. If it does appeal to you, black in between the dotted lines under the column with the heading A. A stands for appeal. If it does not, black it in under heading D. Just go on the basis of your first impression, without spending much time thinking about any one item. Put your name in this space before you begin.

After a patient had completed this test the examiner put before him an MMPI booklet and answer sheet saying:

> This is the last thing which we will do this afternoon. This booklet contains a number of statements and I want you to indicate whether you believe them to be true or false of you by making a mark under the column headed T if true, F if false.

At the end of the afternoon session, the patients were requested to return the following morning. At that time the examiner again distributed the Q sort material and said:

> You remember that the first thing we did yesterday afternoon was to sort these cards into different piles according to whether the statements on them applied to you or didn't. This morning I would like you to sort them again but on a different principle. Instead of sorting them according to the way you are, you should sort them according to the way you would like to be. This means that, in the first step, you should pick out the six statements that you would most like to be like, that you would most like to be true of you, regardless of whether they are actually true of you or not, and the six that you would least like to be true of you regardless of whether they are true of you or not. For example, let's assume that one of the statements is, "I have a million dollars." Yesterday you would have put that in one of the least-like piles since that isn't true of you. Today you might possibly put it in one of the most-like piles since, although it isn't true of you, you might wish that it were true of you. Now, are you all clear on this difference between today's and yesterday's sort? Again, you are to go through the whole stack each time you make a sort.

After completion of the initial step the patients were asked to sort 9, 15, and 20 statements each into most-like and least-like piles. The procedure was the same as in the initial sort.

The examiner then distributed the TAT material and said:

> This is the last thing you will be asked to do. I have given you each a series of pictures which show one or more persons in some kind of a setting. What you are to do in this is to write a story about each picture, a story which has in it the answers to these questions: What might be happening? What might be going on? What might the people be thinking and feeling? What might have led up to this? What might the outcome be? You should not spend more than five minutes on any one story after you have thought of something. Please put your name on each sheet of paper you use.

After a patient had completed the TAT he was dismissed.

The instructions for the second administration varied, depending upon whether the group to be examined consisted entirely of prospective discharges or whether it included "failures" who were about to be transferred to other wards. In the former case the testing session was introduced as follows:

> You may remember that when you first came to 19A you came here and were given several tests. These were part of the routine procedure connected with transfer to that ward. Now you are about to leave the hospital and you will be asked to take these tests again. They are part of the routine procedure connected with the discharge process. Again I would like you to cooperate to the best of your ability.

If the group included "failures" the instructions were:

> When you first came to 19A you came here and were given several tests. These were part of the routine procedure connected with transfer to that ward. It is part of our procedure also, that these tests be taken twice, once upon entering the ward and, the second time, after you have been on the ward a certain length of time. This will be the last time you will be asked to take them. Again I would like you to cooperate to the best of your ability.

In other respects the procedure was the same as that in the initial administrations.

All initial and terminal test data were filed following their completion. The experimenters did not see the data until after the last subject had completed the program.

Attrition Group

During the 24-month period covering the course of the experiment, 25 patients were dropped from the experimental program. These failed to complete the program for one of the following reasons: (*a*) required major surgery; (*b*) discharged with less than 21 days in the experimental program; (*c*) left the hospital without permission; (*d*) developed more acute psychotic symptoms requiring transfer to a closed ward; (*e*) refused to cooperate in posttesting. Table 3 summarizes the reasons for attrition by experimental group.

TABLE 3

Reasons for Attrition by Experimental Group

	GG	G	I	C
Required major surgery		1		2
Discharged in less than 21 days	4		1	2
Developed more acute psychotic symptoms	1		2	3
Left without permission	2	2	1	1
Refused posttesting		1		2
Total	7	4	4	10

TABLE 4

Mean Days in Treatment

	GG	G	I	C	Mean
N	119.5	65.4	117.8	86.0	97.18
Ps	78.6	61.4	129.8	89.5	89.83
Pl	172.3	163.9	191.1	92.5	154.95
Mean	123.47	96.90	146.23	89.33	

Time in Treatment

Since the purpose of this investigation is to determine the efficacy and economy of the various treatment programs for different diagnostic groups, the amount of time spent in treatment, an important consideration in evaluating the efficacy and economy of treatment, was allowed to vary. Table 4 presents the mean time, in days, for the different groups and Table 5 shows the analysis of variance for the same data. Table 5 indicates that significant differences obtain for treatments and diagnoses but not for their interaction.

Table 4 reveals that patients who received individual psychotherapy remained in the hospital longer than the other three treatment groups. A further glance at Table 4 shows that the highly significant difference

TABLE 5

Analysis of Variance of Time in Treatment

Source of Variation	*df*	Mean Square	*F*
Treatments	3	16,228.37	3.98*
Diagnoses	2	40,714.00	9.99**
Interaction:			
Treatment × Diagnoses	6	5,451.67	1.34
Within groups	84	4,073.69	
Total	95		

* Significant at .05 level.

** Significant at .001 level.

between diagnostic groups results from the much longer treatment period of the long-term psychotic group when compared with the other two diagnostic groups. Table 4 also reveals that the long-term psychotic control group remained in treatment a shorter period of time than long-term psychotics in psychotherapy.

RESULTS

The results for each experimental instrument are presented in the following sections.[9] Change is evaluated by: (*a*) comparison of change score[10] variances for the 12 experimental groups and (*b*) comparison of differences between median change scores for treatments, diagnoses, and their interaction.

MMPI

The MMPI, like all experimental instruments, was administered when the patient arrived on the experimental ward and when he left the ward. In addition to the standard scales, the MMPI was scored for Barron's (1953) Ego Strength Scale (*Es*), Welsh's (1956) Anxiety (*A*) and Repression (*R*) scales, and Simon's (1952) Scale of Inner Maladjustment (*SIM*).

A comparison of the change score variances for the experimental groups on each scale was accomplished by use of Bartlett's test (Edwards, 1950).[11]

Heterogeneity of variance obtained on the *F, K, Mf, Pa, Pt, Sc, Es, SIM, R,* and *A* scales. Table 6 presents the mean variances for the 12 groups as well as the treatment and diagnostic groups.

A comparison of the mean variance for treatments, diagnoses, and their interaction is contained in Table 7.

TABLE 6

MEAN VARIANCES FOR THE 12 EXPERIMENTAL GROUPS ON CHANGE SCORES FOR THE 16 MMPI SCALES

	GG	G	I	C	Mean
N	148.56	74.66	128.56	91.08	110.72
Ps	152.89	281.87	216.68	65.89	179.30
Pl	351.76	188.78	189.72	194.88	231.28
Mean	217.74	181.77	178.32	117.28	

Table 7 shows that the mean variances for the 16 MMPI scales are significantly different between treatments, diagnoses, and their interaction. Table 6 reveals that all treatment groups have larger mean variance than the control with the group living treatment showing the greatest mean variance. For the diagnostic groups, short-term psychotics exhibit greater mean variance than nonpsychotics and the long-term psychotic group shows the greatest variance. The higher variances in the short-term psychotic treatment groups compared with controls seems to account for the majority of the interaction variance.

Table 8 presents a comparison of median change scores by treatments, diagnoses, and their interaction. The treatment group comparisons and the interaction effects were evaluated by Friedman's test (Wilcoxon, 1949). The diagnostic group comparisons were made by the median test (Siegel, 1956). In these analyses, the only significant directional change between treatment groups occurred on the *Ma* scale. Here *GG*

TABLE 7

COMPARISON OF TREATMENTS AND DIAGNOSES ON 16 MMPI SCALE VARIANCES

Between Treatments χ_r^2 (3 *df*)	Between Diagnoses χ^2 (2 *df*)	Interaction: Treatment × Diagnoses χ_r^2 (6 *df*)
10.83*	15.15***	19.38**

* Significant at the .05 level.
** Significant at the .01 level.
*** Significant at the .001 level.

---

[9] There were no significant differences (.05 level) among the 12 experimental groups at initial testing on any of the assessment devices.

[10] Change scores refer to initial minus terminal test scores. Further use of the term "change score" refers to this difference.

[11] The change score variances for the 12 experimental groups, together with their tests of significance, on each scale presented in this study may be obtained from the authors.

and I groups showed gains in *Ma,* G group small loss, and C group a large loss. The only significant change between diagnoses occurred on Welsh's (1956) *R* scale. A reduction in *R* scores (repression) between initial and terminal testing occurred in the neurotic and short-term psychotic group, whereas an increase occurred in the long-term psychotic group.

Change scores on the *K, Pd,* and *SIM* scales yielded significant interactions. In all three significant interactions, the greatest share of the variance is attributable to the differential response of long-term psychotics in psychotherapy when compared with the controls. In these three interactions, one or all of the three long-term psychotic groups in psychotherapy demonstrated changes in the maladaptive scale direction while the controls remain relatively the same or change in the adaptive direction. On the other hand, nonpsychotics and short-term psychotics in psychotherapy change in an adaptive scale direction while the same diagnostic groups in the control condition changed only slightly or not at all.

Ward Behavior

Ward behavior is another area of adjustment often used as a treatment criterion. Because of its frequent use and because changes in ward behavior are often said to occur concomitantly with treatment, a Ward Behavior Scale pertinent to the experimental setting was constructed.

The scale was constructed in the following manner. Psychologists, social service personnel, nurses, and aides familiar with the experimental ward setting submitted behavior items which they believed were indicative of patients' adaptation to the ward.

Seventy-six of the 120 submitted items were selected by the experimenters for the Behavior Rating Scale. A review of the literature concerning behavior ratings revealed that 35 of the 76 selected items were so similar to Hospital Adjustment Scale (HAS) items (Ferguson, McReynolds, & Ballachey, 1953) that these 35 HAS items and their alternatives described below, were utilized as part of the Behavior Scale.

The 76 items were then grouped, a priori, into three areas of adaptation. They were (*a*) interpersonal communication—30 items, (*b*) self-care—31 items, and (*c*) participation in work and recreational activities—15 items.

The ratings were keyed in the following manner: for each of the 76 items, the experimenters

TABLE 8

COMPARISON OF TREATMENTS AND DIAGNOSES ON MMPI CHANGE SCORES

MMPI Scale	Between Treatments $\chi_r^2(3\ df)$	Between Diagnoses $\chi^2(2\ df)$	Interaction: Diagnoses × Treatment $\chi_r^2(6\ df)$
1. *Lie*	3.23	3.53	10.87
2. *F*	1.21	0.08	4.25
3. *K*	0.64	0.75	18.75**
4. *Hs*	5.41	0.65	5.56
5. *D*	0.09	1.75	8.00
6. *Hy*	1.10	1.75	8.75
7. *Pd*	3.14	4.09	13.00*
8. *Mf*	2.43	0.59	10.06
9. *Pa*	2.34	1.62	4.81
10. *Pt*	2.34	3.59	8.50
11. *Sc*	1.51	0.75	4.75
12. *Ma*	10.21*	3.05	6.06
13. *Es*	0.14	1.00	10.25
14. *A*	0.94	1.09	8.00
15. *R*	1.06	12.19**	7.56
16. *SIM*	0.65	1.75	16.88**

* Significant at .05 level.
** Significant at .01 level.

made an a priori judgment as to whether the behavior described was adaptive or maladaptive in the ward setting. Then, four raters (one nurse and three aides) rated every patient one week after his admission to the experimental ward and again during his final week on the ward. Each item had two alternatives, one adaptive, the other maladaptive, i.e., "The patient usually will reply if you speak to him," or, "The patient usually does not reply if you speak to him." An item was scored in the following manner. If the four raters checked the adaptive alternative, the item was scored four; if three of the four raters checked the adaptive alternative, the item was scored three; if two raters checked the adaptive alternative, the score was two; if one rater checked the adaptive alternative, the score was one; if no rater checked the adaptive alternative, i.e., if all four checked the maladaptive, the score for the item was zero. A patient's total score was the sum of all items. His score for the three adjustment areas was the sum of the items for each area.

Comparison of the change score variances for the 12 experimental groups is presented in Table 9. Bartlett's (Edwards, 1950) test yields a χ^2 of 44.76 which, for 11 *df*, is significant beyond the .001 level. The marginal totals in Table 9 reveal that the greatest absolute behavioral change occurred in the group living therapy situation and, diagnostically, in the long-term psychotic groups.

A comparison of behavior change medians is presented in Table 10.

This table reveals that no significant differences in ward behavior median change scores obtain between treatments, diagnoses, or their interaction.

TABLE 9

VARIANCES FOR THE 12 EXPERIMENTAL GROUPS ON WARD BEHAVIOR CHANGE

	GG	G	I	C	Mean
N	1,157.00	757.43	319.86	1,108.14	835.61
Ps	888.57	188.29	828.00	399.71	576.14
Pl	7,713.42	4,101.00	1,601.42	1,105.00	3,630.21
Mean	3,253.00	1,682.24	916.43	870.95	

TABLE 10

COMPARISON OF TREATMENTS AND DIAGNOSES ON BEHAVIOR CHANGE

Ward Behavior Scale	Between Treatments $\chi_r^2(3\ df)$	Between Diagnoses $\chi^2(2\ df)$	Interaction: Treatment × Diagnoses $\chi_r^2(6\ df)$
Total Behavior Rating	.86	1.00	4.25
Personal Care	4.19	0.08	4.25
Participation in Activities	1.00	3.41	1.25
Interpersonal Communication	.86	.58	5.56

Holland Vocational Preference Inventory (HVPI)

It has been found by Holland (1958) and others that vocational preferences reflect personality characteristics. Changes in life adjustment could, theoretically, be revealed by such an inventory. The scale selected to evaluate this area was the Holland Vocational Preference Inventory (1958).

Holland's Inventory contains 300 items comprising 12 scales. These scales are briefly described as follows:

Scale 1. Physical Activity. This scale represents a concern with motor skills, physical strength, and a concrete approach to life problems.

Scale 2. Intellectuality. This scale measures preferences for conceptual methods of adjustment and strong needs for understanding the world and perhaps one's self. High scores may represent healthy ego functioning ("ego strength").

Scale 3. Responsibility. This scale appears to be an "oral receptive" scale. High scorers are dependent, verbal, and responsible.

Scale 4. Conformity. This scale is a measure of extraception in the sense of conforming, status-oriented, and conventional behavior.

Scale 5. Verbal Activity. This is an "oral aggressive" scale. High scorers are characterized as aggressive, oral, concerned with leadership and power.

Scale 6. Emotionality. This scale assesses the stability of the subject. High scorers are characterized as emotional and expressive. Low scorers are characterized as cool or "flattened."

Scale 7. Control. This scale measures the habitual inhibition of impulses to act out motives. High scorers are regarded as "over-controlled," and low scorers as "impulsive and under-controlled."

Scale 8. Aggressiveness. This scale is a measure of aggressiveness with high scores representing aggressiveness and low scores passivity.

Scale 9. MF. This is an empirically constructed masculinity—femininity scale. High scores represent identification with or choice of typically masculine occupational roles; low scores represent the choice of feminine occupational roles.

Scale 10. Status. This scale represents the subject's need for status or prestige.

Scale 11. Infrequency. This scale is similar to the MMPI *F* scale. It is intended to provide an estimate of the degree to which the subject makes unusual or infrequent occupational choices.

Scale 12. Acquiescence. This scale indicates the response set of the subject, or the degree to which he responds "like."

Bartlett's test (Edwards, 1950) reveals that significantly different variances occur on the following scales: Conformity, Oral Aggression, Emotionality, *MF*, and Acquiescence.[12] The mean variances for the 12 experimental groups are presented in Table 11. Table 12 shows significant differences between mean variances for treatments and interaction. Table 11 shows the greatest mean variance in the three psychotherapy groups when compared to the control group. This table also reveals that the significant interaction is primarily due to an increase in variance of long-term psychotics in psychotherapy, whereas no such increase appears in the controls.

TABLE 11

MEAN VARIANCES FOR THE 12 EXPERIMENTAL GROUPS ON CHANGE SCORES FOR THE 12 HVPI SCALES

	GG	G	I	C	Mean
N	24.43	24.87	48.67	38.51	34.12
Ps	35.95	33.12	47.50	13.94	32.63
Pl	53.49	82.22	104.78	19.02	64.88
Mean	37.96	46.74	66.98	23.82	

TABLE 12

COMPARISON OF TREATMENTS AND DIAGNOSES ON 12 HVPI SCALE VARIANCES

Between Treatments $\chi_r^2(3\ df)$	Between Diagnoses $\chi^2(2\ df)$	Interaction: Treatment × Diagnoses $\chi_r^2(6\ df)$
30.25**	4.16	20.52*

*.01 level of significance.
**.001 level of significance.

Comparisons of treatments and diagnoses on median change are presented in Table 13.

Table 13 indicates that there are no significant differences between treatments or diagnoses. Two significant interactions occur—one on the intellectuality scale and the other on the masculinity–femininity scale. These two significant interactions, however, do not show any consistent differences when the median change scores for each of the 12 experimental groups are investigated. Due to the number of tests of significance

TABLE 13

COMPARISON OF TREATMENTS AND DIAGNOSES ON HVPI CHANGE SCORES

HVPI Scale	Between Treatments $\chi_r^2(3\ df)$	Between Diagnoses $\chi^2(2\ df)$	Interaction: Treatment × Diagnoses $\chi_r^2(6\ df)$
1. Physical Activity	3.24	2.95	7.75
2. Intellectuality	1.71	.75	15.13*
3. Responsibility	2.71	2.18	9.32
4. Conformity	5.26	.25	5.26
5. Verbal Activity	3.35	1.62	10.25
6. Emotionality	1.69	.75	6.75
7. Control	5.08	.25	3.31
8. Aggressiveness	1.18	1.02	3.06
9. Masculinity-Femininity	5.09	.61	17.75**
10. Status	5.85	1.11	5.19
11. Infrequency	.13	.33	2.06
12. Acquiescence	2.65	1.09	7.63

* Significant at the .05 level.
** Significant at the .01 level.

---

[12] Difference scores were computed in two ways. On those scales where the highest score is considered most adaptive, the difference scores were determined by substracting the initial from the terminal test score. On other scales, where the mean scale score is considered most adaptive, difference scores were determined by computing the unit movement toward or away from the mean.

computed, one should approach these findings with caution.

Q Sort

Q sort is another instrument frequently utilized to evaluate changes that occur during psychotherapy. In this study, a 100-item sort with five scales of 20 items each was used.

The general factor structure of Sheer's[13] Sort was followed; however, items appropriate to the experimental population were constructed. The five areas that the *Q* sort attempted to cover were: (*a*) social adaptability, (*b*) emotional control, (*c*) conformity, (*d*) self-acceptance, and (*e*) self-confidence. These were sorted both for self and ideal on initial and terminal testing. Each of the five scales contained 10 negative and 10 positive statements, the directionality of a statement being determined by the a priori judgment of the authors. For example, the self-confidence scale had as a positive item, "I can argue for my point of view" and, as a negative item, "I do as I am told." In scoring, positive items were given the score of the stack into which they were sorted,[14] and negative items received

TABLE 14

CORRELATIONS BETWEEN SELF AND IDEAL *Q* SORTS AT INITIAL AND TERMINAL TESTING

Subject	GG		G		I		C	
	Initial	Terminal	Initial	Terminal	Initial	Terminal	Initial	Terminal
	r	*r*	*r*	*r*	*r*	*r*	*r*	*r*
Nonpsychotic								
1	.67	.39	.45	.62	.53	.71	.32	.37
2	−.14	.39	.39	.42	.10	−.33	−.02	.03
3	−.35	−.12	−.51	−.43	.36	.60	.81	.79
4	.39	.61	.46	.42	.34	.49	.10	−.02
5	−.17	−.20	.78	.90	.14	.66	.64	.47
6	.52	.66	.28	.09	−.45	−.49	−.07	.09
7	.03	−.02	.54	.43	.18	−.27	.19	.58
8	−.48	−.31	−.25	.24	−.07	.15	.50	.34
Short-term Psychotic								
1	.34	.32	.21	.76	.50	.60	.09	−.19
2	.41	.33	−.14	.50	−.08	−.03	.06	.22
3	.66	.76	−.07	.09	.04	.05	.58	.64
4	−.11	.37	−.14	.02	.58	.73	.65	.65
5	.38	.68	.54	.61	.57	.69	.50	.46
6	−.09	.33	−.04	.21	−.18	.77	.73	.81
7	−.07	.00	.11	−.02	.56	.63	.51	.61
8	.46	.41	.11	.19	−.54	.65	.74	.81
Long-term Psychotic								
1	.32	−.08	.27	.19	−.16	.14	.78	.52
2	−.04	.37	−.08	.72	−.25	−.33	.58	.45
3	−.15	.43	−.05	.38	.25	−.23	.61	.58
4	.22	−.15	.53	.59	−.16	−.48	−.52	−.09
5	.23	.26	.12	.34	−.72	−.66	.07	.42
6	−.67	−.38	.65	.45	.58	.48	.87	.83
7	.31	−.08	.60	.29	.28	.54	.53	.61
8	.05	.10	.59	.24	.58	.03	.30	−.02

---

[13] D. E. Sheer. A self-concept test: Reliability and validity data. Personal communication. 1955.

[14] For instructions and pile numbers, the reader is referred to the section on Experimental Instruments and Testing Situation.

TABLE 15

VARIANCES FOR 12 EXPERIMENTAL GROUPS ON z SCORE CHANGE

	GG	G	I	C	Mean
N	.069	.074	.151	.064	.090
Ps	.069	.119	.286	.036	.128
Pl	.169	.229	.094	.117	.152
Mean	.102	.141	.177	.072	

the score of nine minus the pile number. The total score for each factor was the arithmetic sum of the 20 items in the scale. Thus, the following sets of measurements were obtained from the Q sort: (*a*) correlations between the 100-item self and ideal sorts at initial and terminal testing; (*b*) scores on self-sort on the five aforementioned scales at initial and terminal testing; (*c*) score on ideal-sort on the five scales at initial and terminal testing.

The usual comparisons made on Q sort data involve correlations between self- and ideal-sort at initial testing and between the same measures at terminal testing. Theoretically, successful therapy is evidenced by a positive increase in the correlations.

The correlations between self- and ideal-sorts at initial and terminal testing are presented in Table 14. These correlations were then transformed into z scores, and a population of differences between z scores (initial-terminal) was obtained.

The variances of this distribution for the 12 experimental groups is presented in Table 15. Barlett's test (Edwards, 1950) yields a χ^2 of 13.42 which, for 11 *df*, does not reach the .05 level of significance; however, the trend of larger variances in the three psychotherapy treatments and long-term psychotic diagnostic group occurs again.

TABLE 16

COMPARISON OF TREATMENTS AND DIAGNOSES ON z SCORE CHANGE

Between Treatments	Between Diagnoses	Interaction: Treatment × Diagnoses
χ_r^2(3 *df*)	χ^2(2 *df*)	χ_r^2(6 *df*)
2.85	4.00	4.50

TABLE 17

MEAN VARIANCES FOR THE 12 EXPERIMENTAL GROUPS ON CHANGE SCORES FOR THE 5 SELF-SORT SCALES

	GG	G	I	C	Mean
N	113.11	124.80	140.26	151.20	132.34
Ps	150.60	340.06	840.63	72.77	351.02
Pl	311.43	526.95	145.23	126.09	277.43
Mean	191.71	330.60	375.37	116.69	

Table 16 presents a comparison of treatments, diagnoses, and their interaction on the median differences between z scores. This table reveals that no significant differences obtain.

The previous comparisons dealt with self-ideal correlations based upon the 100 items. The following analyses are for the five Q sort scales of 20 items each. The first analysis compares change scores on self-sort alone. A Bartlett's test (Edwards, 1950) comparison of the self-sort change score variances showed significant differences between the variances on the social adaptability, self-acceptance, self-confidence, and emotional control scales.

The average variances for the five scales are presented in Table 17. A nonparametric analysis of the variances for the five scales is presented in Table 18 which shows that

TABLE 18

COMPARISON OF TREATMENTS AND DIAGNOSES ON 5 SELF-SORT SCALE VARIANCES

Between Treatments	Between Diagnoses	Interaction: Treatments × Diagnoses
χ_r^2(3 *df*)	χ^2(2 *df*)	χ_r^2(6 *df*)
4.28	5.20	22.80*

* Significant at the .001 level.

TABLE 19

COMPARISON OF TREATMENTS AND DIAGNOSES ON SELF-SORT CHANGE SCORES

Self-Sort Scale	Between Treatments χ_r^2 (3 df)	Between Diagnoses χ^2(2 df)	Interaction: Treatment × Diagnoses χ_r^2(6 df)
1. Emotional Control	4.85	.58	5.25
2. Conformity	.71	5.25	2.25
3. Self-Acceptance	3.18	2.58	15.75*
4. Social Adaptability	3.03	1.59	15.56*
5. Self-Confidence	3.29	.75	18.87**

* Significant at the .05 level.
** Significant at the .01 level.

the interaction alone is significant. Table 17 suggests that the interaction is significant because nonpsychotics show relatively equal variances over the four treatments, short-term psychotics show larger variances in the three groups receiving psychotherapy with the largest in the individual therapy group, and long-term psychotics show the largest variances in the group living and group therapy treatments.

Table 19 presents a comparison of the medians for the self-sort change scores. No significant differences obtain on any of the five scales between treatments or diagnoses. However, three significant interactions obtain.

When the three significant interactions are viewed collectively, it is evident that the major contributor to the significant interaction is the differential response of the diagnostic groups to all three psychotherapy treatments compared with the control. The following schema generally holds for the three scales:

	All Therapies	Control
Nonpsychotics	Large Positive Change	No Change or Small Positive Change
Short-term Psychotics	Large Positive Change	Small Positive or Small Negative Change
Long-term Psychotics	Moderate Negative Change	Large Positive Change

TABLE 20

MEAN VARIANCES FOR THE 12 EXPERIMENTAL GROUPS ON CHANGE SCORES FOR THE 5 IDEAL-SORT SCALES

	GG	G	I	C	Mean
N	86.94	47.37	138.42	140.77	103.38
Ps	95.80	85.67	237.24	49.16	116.97
Pl	200.91	174.72	162.95	95.55	158.53
Mean	127.88	102.59	179.54	95.16	

As was previously mentioned, initial and terminal ideal-sorts were also obtained. These were scored on the five scales.

A comparison of the ideal-sort change score variances for the 12 experimental groups revealed significantly different variances on four of the five scales: emotional control, conformity, self-acceptance, and self-confidence.

The average of the five scale variances for the 12 experimental groups is shown in Table 20. An analysis of the five variances when treated as scores appears in Table 21. Only significant differences between the variances of the treatment groups obtain. Table 20 indicates that the variances in descending order of magnitude are: individual therapy, group living, group therapy, and control.

An analysis of the median change scores on ideal-sort is presented in Table 22. Significantly different changes occurred between diagnostic groups on the emotional control, conformity, and social adaptability scales. The short-term psychotic group has the largest median improvement in emo-

TABLE 21

COMPARISON OF TREATMENTS AND DIAGNOSES ON THE 5 IDEAL-SORT SCALE VARIANCES

Between Treatments χ_r^2(3 df)	Between Diagnoses χ^2(2 df)	Interaction: Treatments × Diagnoses χ_r^2(6 df)
9.66*	2.80	10.40

* Significant at the .05 level.

TABLE 22

Comparison of Treatments and Diagnoses on Ideal-Sort Change Scores

Ideal-Sort Scale	Between Treatments $\chi_r^2(3\ df)$	Between Diagnoses $\chi^2(2\ df)$	Interaction: Treatment × Diagnoses $\chi_r^2(6\ df)$
1. Emotional Control	2.60	6.59*	3.94
2. Conformity	7.00	7.60*	5.50
3. Self-Acceptance	2.88	1.58	5.25
4. Social Adaptability	2.16	6.78*	5.06
5. Self-Confidence	1.24	1.59	8.75

*.05 level of significance.

tional control and conformity, whereas both the non- and short-term psychotic groups demonstrate greater median improvement in the area of social adaptability than do the long-term psychotics.

Thematic Apperception Test

Eight selected TAT cards were presented to the subjects at the initial and terminal testing sessions. The cards were selected by the authors for their value in evoking responses concerning interpersonal relationships. In order of presentation they were: 1, 2, 4, 9, 13 MF, 7 BM, 6 BM, and 12 M.

Protocols were scored in the following manner: All identifying information was first removed. Then, the initial and final protocols for each patient were presented to the experimenter who made a judgment as to whether one protocol was more adaptive or whether they were adaptively equivalent. The scoring system for the protocols was:

I. Protocol A is judged more adaptive than Protocol B when Protocol B has four or more rejected cards and Protocol A has none.

II. Protocol A is judged more adaptive than Protocol B when A's stories reveal developed plots and B's are descriptions of the cards.

III. Protocol A is judged more adaptive than B if both reveal plots but A's are more well-developed than B's.

IV. Protocol A is judged more adaptive than B when both reveal well-developed plots but A's stories contain direct affective statements, i.e., love, hate, etc., while B's do not.

V. Protocol A is judged more adaptive than B when both reveal well-developed plots, both contain direct affective statements, but Protocol A has more positive statements concerning interpersonal relationships than B.

VI. Protocol A is judged more adaptive than B when both reveal well-developed plots, both contain direct affective statements, both have an equivalent number of positive statements concerning interpersonal relationships, but A's outcome is more positive than B's outcome.

After the judgments were completed, identifying information was revealed and scores were recorded. The scores were recorded by comparing final with initial protocols as follows: final more adaptive than initial, adaptively equivalent, or less adaptive than initial.

Two experimenters independently compared the initial and final protocols of the same 20 patients by this scoring system. They agreed in 18 of the 20 (or 90%) of the cases.

For statistical analysis the scores were dichotomized into protocols showing posi-

TABLE 23

TAT Changes for the 12 Experimental Conditions by %

		GG	G	I	C	Total %
N	Positive change	75.0	50.0	12.5	62.5	50.0
	No change and negative change	25.0	50.0	87.5	37.5	50.0
Ps	Positive change	37.5	75.0	37.5	0.0	37.5
	No change and negative change	62.5	25.0	62.5	100.0	62.5
Pl	Positive change	25.0	50.0	75.0	50.0	50.0
	No change and negative change	75.0	50.0	25.0	50.0	50.0
	Total %					
	Positive change	45.8	58.3	41.7	37.5	
	No change and negative change	54.2	41.7	58.3	62.5	

tive change and those showing no change or negative change. Table 23 presents the percentage of cases in the two categories for each of the 12 cells.

Sheffield's (1957) method was used to compare the treatment and diagnostic groups. Table 24 presents the analysis.[15]

Table 24 indicates that differences between treatments and diagnoses do not reach the .05 level of significance. The interaction is significant at the .01 level. Reference to Table 23, however, suggests that this interaction is significant because more positive TAT change is brought about by group methods and control than with individual therapy when nonpsychotics are treated; more positive TAT change occurs with the three psychotherapy groups than controls when short-term psychotics are treated; and more positive TAT change is brought about by individual therapy when long-term psychotics are treated.

TABLE 24

COMPARISON OF TREATMENTS AND DIAGNOSES ON TAT CHANGE

	Squares	*df*	*MS*	*F*
Treatments	4.67	3	1.56	0.53
Diagnoses	2.67	2	1.34	0.67
Interaction: Treatment × Diagnoses	35.33	6	6.06	3.06**
Total	42.67	11	3.88	1.96*

*.05 level of significance.
**.01 level of significance.

Follow-up

A follow-up questionnaire was obtained for each patient six months after he had left the experimental program. The questionnaire was completed by an individual with whom the patient was living, or, if he had not left or had returned to the hospital, by a psychologist familiar with the case.[16]

Follow-up Scale. The questionnaire was constructed to encompass several areas of life adjustment. Simplicity of items was stressed. The items were multiple choice, with a designated space for remarks if the respondent believed the structured choices to be inadequate. The questionnaire consisted of five objective and four subjective items. The five objective items concerned: (*a*) percentage of time employed, (*b*) hospitalization, (*c*) alcoholic indulgence, (*d*) antisocial behavior, and (*e*) number of friends. The four subjective items involved a judgment by the respondent concerning patient's (*a*) amount of verbal communication with others, (*b*) general adjustment, (*c*) degree of problem behavior, and (*d*) degree of illness.

For analysis, the nine items were divided into the following dichotomies:

1. Was, or was not, rehospitalized
2. Was, or was not, employed full-time
3. Has some, or no, close friends
4. Frequently, or rarely, talks to the people with whom he lives
5. Getting along satisfactorily, or poorly
6. Usually hard, or easy, to get along with
7. Well, or sick, most of the time
8. Is, or is not, a problem drinker
9. Has, or has not, been in trouble with the law

Each of the nine items, therefore, has an arbitrary adaptive and maladaptive alternative. The tetrachoric correlations among the nine follow-up items are presented in Table 25.

Table 25 reveals that, in general, positive relationships exist among the items. The high degree of relationship among Items 5, 6, and 7 indicates a common parameter and the theoretical possibility of item substitution. However, the remaining relationships, while generally positive and often significant, are rarely of sufficient magnitude to warrant the substitution of one item for another.

Comparisons of the follow-up items by treatments, diagnoses, and their interactions, computed by Sheffield's (1957) method, are presented in Table 26.

Table 26 indicates that the amount of employment item is the only one that significantly differentiates the treatment groups. The percentage of patients from each treatment group employed full-time is as fol-

---

[15] The within-cell variance is 1.98.

[16] Information was obtained on all 96 patients for Items 1 and 2. On Items 3 through 9, information was obtained on 94 of the 96 subjects. The scores of the two missing patients was the modal score for his experimental group; however, deletion of these two patients would not alter the results. Eleven of the 96 patients were failures, i.e., did not leave the hospital. Viewed diagnostically, nine were in the long-term psychotic group and two in the nonpsychotic group. Viewed by treatments, four were in group living (GG), two in group therapy (G), and five in individual therapy (I).

TABLE 25

Correlations among the 9 Follow-up Items

	1	2	3	4	5	6	7	8	9
1. Rehospitalization									
2. Full-time employment	.66**								
3. Number of friendships	.12	.46*							
4. Amount of interpersonal communication	.35	.45	.78**						
5. General adjustment	.56*	.40	.54*	.55*					
6. Degree of problem behavior	.42	.42	.78**	.36	.90**				
7. Degree of illness	.59**	.57*	.50*	.36	.93**	.91**			
8. Type of alcoholic indulgence	.29	.14	−.08	.11	.77**	.59*	.60*		
9. Legal violations	.65**	.25	−.22	−.22	.71*	.62*	.44	.75**	

* Significant at .01 level.
** Significant at .001 level.

lows: Group living, 42%; Group therapy, 33%; Individual therapy, 42%; and Control, 8%. It can readily be seen that the three treatments involving psychotherapy have a higher percentage of employed than the control and that the group living and individual therapy treatments have the highest percentage of the three treatments. Table 1 indicates that this result cannot be explained by differences in employment prior to the study.

Diagnostically, six of the nine items significantly differentiate the groups. Table 27 presents the percentage of the three diagnostic groups in the adaptive alternative for each item. This table indicates that in the areas of hospitalization, employment, and friendships, the short-term psychotic group displayed the most adequate adjustment, the nonpsychotics next, and the long-term psychotics the poorest adjustment. In the two areas involving judgments concerning

TABLE 26

Comparison of Experimental Groups on Follow-up Items

Follow-Up Area	Between Treatments *F* ratio (3 *df*)	Between Diagnoses *F* ratio (2 *df*)	Interaction: Treatment × Diagnoses *F* ratio (6 *df*)
1. Rehospitalization	0.16	8.16**	1.02
2. Amount of employment	2.78*	6.25**	0.50
3. Number of friendships	1.26	4.81**	1.60
4. Amount of interpersonal communication	0.56	2.18	1.07
5. General adjustment	0.55	3.05*	0.44
6. Degree of problem behavior	2.08	0.89	0.78
7. Degree of illness	0.66	3.89*	0.81
8. Type of alcoholic indulgence	0.83	4.56*	1.33
9. Legal violations	0.15	2.12	0.92

Note.—The within-cell variance is 1.98.
* Significant at the .05 level.
** Significant at the .01 level

TABLE 27

% OF DIAGNOSTIC GROUPS SHOWING BEST ADAPTATION ON THE 6 SIGNIFICANT FOLLOW-UP ITEMS

Follow-up Area	Non-Psychotic	Short-Term Psychotic	Long-Term Psychotic
1. Out of hospital	50.0	87.5	40.6
2. Employed full-time	34.4	50.0	9.4
3. Some friends	62.5	75.0	37.5
5. Getting along satisfactorily	53.1	81.3	59.4
7. Well most of the time	53.1	84.4	59.4
8. Not a problem drinker	65.6	90.6	90.6

general adjustment (Items 5 and 7), again the short-term psychotics were rated as making a significantly better adjustment than either the nonpsychotics or long-term psychotics. Problem drinkers are more prevalent in the nonpsychotic group.

Relationships among Selected Psychotherapeutic Criteria

Thirteen scales chosen because they are frequently used as psychotherapeutic criteria and because they measure adjustment in different areas were selected for correlational analysis. These scales were: follow-up measures of Rehospitalization, Employment, and Degree of Illness; *Sc, Es, SIM, A,* and *R* scales of the MMPI; Q sort; TAT; Ward Behavior; Job Behavior;[17] and Finney's (1954) Group Therapy Scale.

Only the last two scales will be described here since the others were discussed in preceding sections. The Job Behavior Scale included 31 items dealing with the patient's performance and interpersonal adjustment in his hospital job assignment. These items, like those in the Ward Behavior Scale, were stated in terms of adaptive and maladaptive alternatives; i.e., "The patient usually profits by his mistakes" or "The patient repeats his mistakes over and over again." The rating was completed by each patient's job supervisor in the first and final week of the experimental program. A reliability coefficient of .93 was found for 20 patients, by the method of successive ratings. Each patient in group therapy was rated on the Group Therapy Scale every 30 days.

All scales were scored so that the most adaptive scores were numerically highest and the least adaptive were lowest. Thus, only positive correlations indicate an association between the theoretical adaptive levels of the scales. For all within-hospital measures, final scores completed shortly before discharge were correlated. In the case of the Group Therapy Rating Scale, the last rating was used. For the Q sort, the final correlations between self- and ideal-sort were ranked, with the highest positive correlation receiving the highest rank. The final 96 TATs were divided into two groups —most and least adaptive.[18]

The *N* for all scales was 96 except for Job Behavior with an *N* of 70 and the Therapy Rating Scale which had an *N* of 41. All scales, except the follow-up items, were divided at the median for computational purposes.

The correlations are presented in Table 28. An examination of the matrix reveals two rather independent clusters of items. One involves the three follow-up items and the group therapy behavior rating. The other cluster includes the *Sc, A, Es,* and *SIM* scales of the MMPI, the Q sort, and the TAT as a marginal member.

Three of the scales are not contained in these two clusters. They are Job Performance, Ward Behavior, and Welsh's (1956) *R*. Hospital job behavior is not significantly related to any other measure and Welsh's *R* is correlated only with ward behavior. On the other hand, ward behavior is significantly related only to the *R* scale, group therapy behavior, and TAT.

That test scores are not significantly related to follow-up and are associated with each other is not surprising (Fairweather, Moran, & Morton, 1956). But why is group therapy behavior the only within-hospital measure significantly related to follow-up? The group therapy rating completed by the therapist after the patient's last session was used for the correlational analysis. Thus, this rating involves continuing observations of an individual over several weeks' time. Further, in contrast to ward behavior and job performance ratings, the group therapy ratings were made by psychologists rather than by nurses, aides, and job supervisors. In addition, the group therapy situation involves intense interpersonal experiences which may be a better approximation of the extramural situation than either ward behavior or the hospital job situation. Therefore, both the rater and the situation may have contributed to these significant relationships.

---

[17] The Job Behavior Scale has been deposited with the American Documentation Institute. See Footnote 8.

[18] General principles of scoring outlined in the TAT section were followed.

DISCUSSION

The purpose of this study was to test a series of hypotheses in an attempt to determine the relative efficiency of four commonly utilized therapeutic procedures with three different patient groups. In view of the lack of agreement concerning the validity of a specific criterion for evaluating treatment, several criteria were selected.

In neuropsychiatric hospitals, economic considerations are important in determining the feasibility of therapeutic endeavors because most hospitals operate on limited and frequently minimal budgets. Therefore, the amount of time taken to complete a specified program is a relevant factor in the selection of treatment programs. The analysis of time in treatment yielded significant differences, both for treatment and diagnostic groups. Individual therapy patients remained in treatment longer than the other three groups. Group living took less time, group therapy and control the least, and approximately equivalent, time. By diagnostic group, the long-term psychotic remained in treatment significantly longer than did the short-term and nonpsychotic groups. In terms of time alone, it appears that group therapy and individual work assignment are the most economical. Diagnostically, nonpsychotics or short-term psychotics can be treated most economically (shortest time in hospital).

While time in treatment is an important consideration in the selection of psycho-

TABLE 28

TETRACHORIC CORRELATIONS AMONG THERAPEUTIC CRITERIA

	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Follow-up Rehospitalization													
2. Follow-up Employment	.66[c]												
3. Follow-up Degree of Illness	.59[b]	.57[b]											
4. Finney's Group Therapy Behavior	.42[a]	.75[c]	.09										
5. Ward Behavior	.04	.28	−.32	.50[a]									
6. Job Behavior	.02	−.06	.08	−.14	−.14								
7. MMPI Welsh's *R*	.06	.05	.07	.29	.41[b]	−.24							
8. MMPI *Es*	.30	.15	−.01	.12	.07	−.19	−.10						
9. MMPI Welsh's *A*	.30	.18	.16	.05	.07	.30	−.16	.75[c]					
10. MMPI *Sc*	.21	.02	.29	.11	.23	−.08	.20	.53[b]	.58[c]				
11. MMPI *SIM*	.26	.09	−.09	.12	−.04	−.27	.26	.78[c]	.74[c]	.58[c]			
12. *Q* Sort	.21	.10	−.14	.20	.10	−.29	.07	.64[c]	.75[c]	.66[c]	.87[c]		
13. TAT	.04	.23	−.21	.41[a]	.33[a]	−.27	.10	.13	.20	.26	.16	.47[b]	

[a] .05 level of significance.
[b] .01 level of significance.
[c] .001 level of significance.

therapeutic programs, it sheds little light upon the success or failure of such programs. In an attempt to ascertain the relationship between time in treatment and successful follow-up, a "global follow-up" score[19] was computed for each patient. A comparison of treatments, diagnoses, and their interaction on this criterion is presented in Table 29. The only significant difference is between diagnoses; hence, comparisons of success and time in treatment were independently computed for the three diagnostic groups. Table 30 shows no significant differences in mean time in treatment for the least and most adaptive follow-up groups.[20] Thus, length of time in treatment appears to have little relationship to successful posthospital adjustment. This result is reasonable if we consider psychotherapy as a learning process. Such a process is unique to each individual and each learns at his own rate. If such is the case, the amount learned may be unrelated to time in the learning situation. Perhaps future studies might profitably attempt to discriminate slow and fast learners prior to treatment.

TABLE 29

COMPARISON OF TREATMENTS AND DIAGNOSES ON GLOBAL FOLLOW-UP SCORES

Between Treatments χ_r^2 (3 *df*)	Between Diagnoses χ^2 (2 *df*)	Interaction: Treatment × Diagnoses χ_r^2 (6 *df*)
2.15	21.75*	4.38

* Significant at .001 level.

Change was evaluated in the preceding sections by analyzing the difference scores between initial and final testing on the MMPI, behavior ratings, HVPI, TAT, and Q sort. The first finding was that significant differences in variances obtain between treatment and diagnostic groups. When change scores are viewed globally, the generic conclusion, that the three treatments involving psychotherapy demonstrate significantly more absolute change than the control, seems warranted. These differences do not appear to be simply a function of time between tests. Table 9 shows that the individual treatment group presents approximately the same variance as the controls despite the fact that these two groups spent the greatest and least time in treatment, respectively. Further, absolute change seems to be related to type of therapy involved. For example, the greatest variance in ward behavior is achieved by the group

TABLE 30

COMPARISON OF MEAN DAYS IN TREATMENT FOR PATIENTS IN LEAST AND MOST ADAPTIVE FOLLOW-UP CATEGORIES

	Diagnoses					
	N		Ps		Pl	
	Most Adaptive	Least Adaptive	Most Adaptive	Least Adaptive	Most Adaptive	Least Adaptive
Mean	117.57	81.25	89.35	90.58	160.56	147.71
SD	65.35	45.35	71.18	75.35	74.22	57.55
t	1.79		.04		.54	

---

[19] All nine follow-up items were dichotomized into adaptive and maladaptive alternatives. The adaptive alternative received a score of one and the maladaptive a score of zero. Each patient's score was the sum of the nine items.

[20] Each diagnostic group was divided at its median global follow-up score. Least and most adaptive refer to patients below and above this median.

living subjects, whereas the greatest Q sort variance is achieved by the individual therapy group. Group living therapy was oriented toward interpersonal behavior in a variety of situations, while individual therapy was more closely oriented toward perceptual change. Thus, the different variances of the treatment groups in different areas of change might be, in some measure, an indication of the attendant responses in the learning situation.

While significantly larger variances were found for treatment groups as compared to controls, significantly different variances also occurred between diagnostic groups. Generally, significantly larger variances obtained for the long-term psychotics when compared with the short-term and nonpsychotic groups. Again, it seems highly unlikely that the markedly different variances are attributable to time alone, although clearly they are confounded with time. From observation of the patients in the treatment processes, it seems likely that the significantly greater change score variances of the long-term psychotic group can be attributed, in part, to the greater range of responses available to them. In this regard, the repertoire of responses available to the long-term psychotic in reacting to a close interpersonal relationship ranges from, to take one example, complete withdrawal to hyperactive participation. By comparison, the range of responses available to the nonpsychotic is much more restricted and, on this theoretical continuum, the short-term psychotics would fall somewhere between these two groups. In addition, long term psychotics in the three psychotherapy groups contribute greatly to the total long-term psychotic group variance, while the controls contribute little. Thus, as far as the scales are concerned, the long-term psychotics, and to a lesser but usually significant extent, the short-term psychotics, change in either a positive or negative scale direction in psychotherapy while the same diagnostic groups remain comparatively unchanged without psychotherapy.

There are, of course, several possible explanations for the different variances. One is that patients in therapy get better or worse but generally do not remain the same, whereas patients not in therapy remain about the same. A second is inherent in the construction of the scales themselves. Most clinical instruments are constructed on the assumption of an underlying theoretical dimension of adjustment. Thus, positive change on a scale usually refers to an increment of movement along the scale in the adaptive direction. In this context, subjects show improvement only by moving in the same direction. Clinically, however, this assumption may not be tenable. For example, two patients bearing the same diagnostic label, one a rigid conformist and the other a nonconformist, might be expected, if therapy was successful, to move in opposite directions. One might move 20 positive scale units and the other 20 negative units on a Conformity Scale. The mean of the two changes is zero. Two untreated patients might move two positive and two negative units. Here again, the mean is zero, and the difference between means is obviously not significant. The variances, however, of the two groups would reveal the change. That movement in either direction on a scale may be an adaptive change for the individual patient is emphasized in

TABLE 31

COMPARISON OF SELECTED CHANGE SCORES FOR PATIENTS IN LEAST AND MOST ADAPTIVE FOLLOW-UP CATEGORIES

Scale	Diagnoses		
	N	Ps	Pl
	χ^2	χ^2	χ^2
Job Behavior	.05	.39	.88
Ward Behavior	.13	.00	.00
MMPI *F*	.98	.68	.00
MMPI *K*	.00	.01	2.17
MMPI *Sc*	.00	.00	.45
MMPI *Es*	6.22*	1.20	1.14
MMPI *A*	.00	.01	.04
MMPI *SIM*	.13	.68	.13
HVPI 2	.98	.01	.45
HVPI 3	.02	.01	.00
Q Sort	1.14	.15	.00

* Significant at .05 level.

two additional analyses. Table 31 shows that only one significantly greater positive change occurs for the least and most adaptive follow-up groups, and this change was in the opposite direction predicted, i.e., least adaptive on follow-up showed greater positive scale change. On the other hand, Table 32 reveals significantly different variances for the least and most adaptive follow-up groups on five scales for the short-term psychotics and three for the long-term psychotics. Furthermore, on each significantly different scale, the larger variance was associated with the most adaptive follow-up. Thus, it seems that change in either scale direction can be adaptive for the individual patient. For this reason, it might be clinically more meaningful if a directional prediction of adaptive change were made for each patient prior to therapy. Thus, positive change for a patient would be determined by the direction in which clinicians agree that the patient should move rather than by the scale itself.

Although generally the variances of the change scores significantly differentiated the treatment groups, differences between change score medians did not. Of the within-hospital measures, only the *Ma* scale of the MMPI showed a significant difference and, in view of the large number of measures, it seems plausible that little significance should be attached to this finding.

Interactions, on the other hand, revealed significance on nine scales. On six of the nine significant interactions, the three psychotherapy groups showed change in an adaptive direction for the short-term and nonpsychotic groups, while an adaptive decrement occurred with the long-term psychotics. Generally, nonpsychotics and short-term psychotics in the control group remained relatively unchanged, while long-term psychotics showed a small or large adaptive increment. As far as scale score adaptiveness is concerned, then, nonpsychotics and short-term psychotics change more adaptively with psychotherapy while long-term psychotics change more adaptively without therapy.

On within-hospital measures, the diagnostic groups showed significantly different

TABLE 32

COMPARISON OF SELECTED CHANGE SCORE VARIANCES FOR PATIENTS IN LEAST AND MOST ADAPTIVE FOLLOW-UP CATEGORIES

	Diagnoses								
	N			Ps			Pl		
Scale	Most Adaptive s^2	Least Adaptive s^2	*F*	Most Adaptive s^2	Least Adaptive s^2	*F*	Most Adaptive s^2	Least Adaptive s^2	*F*
Job Behavior	9.80	31.81	3.25	20.25	11.00	1.84	113.61	8.98	12.65*
Ward Behavior	388.25	934.02	2.41	735.30	186.22	3.95*	2218.31	2122.67	1.05
MMPI *F*	85.63	45.14	1.90	197.41	30.91	6.39*	340.17	331.23	1.03
MMPI *K*	32.12	52.69	1.64	74.29	52.06	1.43	254.10	52.39	4.85*
MMPI *Sc*	309.39	145.89	2.12	397.41	369.74	1.28	305.58	635.84	2.09
MMPI *Es*	51.35	36.40	1.41	89.54	32.85	2.73	182.31	47.03	3.88*
MMPI *A*	66.39	53.92	1.23	174.29	16.72	10.42*	167.53	64.23	2.61
HVPI 2	11.67	34.56	2.96	59.15	26.92	2.20	80.11	101.35	1.27
HVPI 3	29.39	37.57	1.28	110.03	26.39	4.17*	115.06	86.29	1.33
MMPI *SIM*	262.37	324.25	1.24	510.95	195.75	2.61	1142.02	549.54	2.08
Q sort	.09	.06	1.50	.16	.04	4.00*	.09	.20	2.22

* Significant at .05 level.

change on four scales. Although such findings must be interpreted cautiously, they suggest, as set forth in the second hypothesis, that the short-term and nonpsychotic groups improve more with treatment than long-term psychotics.

Follow-up items showed significant differences between treatment groups on the employment item. All three treatment groups showed a greater percentage of full-time employed than the controls, which supports the first hypothesis. The group living and individual therapy groups have the highest percentages of employed patients. The eight other items did not significantly differentiate the treatment groups. If employment were chosen as the criterion of posthospital adjustment, then either the group living or group therapy treatments are desirable because they take less time and personnel than individual therapy and are significantly more successful than the control. Further, such therapy would be more efficient with nonpsychotics and short-term psychotics than with long-term psychotics. On the other hand, if any other follow-up item besides employment were chosen as the criterion, e.g., rehospitalization, then hospital work alone would be the most efficient treatment. In this regard, it should be noted that follow-up employment is the only item yielding any measure of the patient's contribution to others. It will be recalled that nine significant interactions occur on the within-hospital measurements, but no significant interaction is found on follow-up. Thus, differential test score change of the diagnostic groups to treatment is not revealed by the follow-up.

Diagnostically, six of the nine items significantly differentiated the groups; the short-term psychotics demonstrate the most adequate global posthospital adjustment, nonpsychotics are next, and long-term psychotics have the poorest adjustment. These findings again support the second hypothesis. The significant differences on follow-up ratings indicate all too clearly the problem facing treatment planners. The long-term psychotic, in this study, is much less responsive to psychotherapy than the nonpsychotic or short-term psychotic.

In addition, Table 33 indicates that patients who did not receive tranquilizing medication did as well on follow-up as patients who did. Further, the table reveals that the results were the same for all three diagnostic groups. Since tranquilizing medication is frequently used to improve ward behavior, i.e., reduce management problems, and since ward behavior is unrelated to follow-up, it is not surprising that the patients who did not receive drugs did as well on follow-up as those who did.

It is interesting to note that nine of the eleven failures in the total program were long-term psychotics in psychotherapy. No failures occurred in the control group. On the other hand, the three full-time employed long-term psychotics were in psychotherapy and none in the control. This finding suggests that long-term psychotics in intensive

TABLE 33

COMPARISON OF PATIENTS WHO RECEIVED TRANQUILIZING MEDICATION IN LEAST AND MOST ADAPTIVE FOLLOW-UP CATEGORIES

Diagnoses								
N			Ps			Pl		
	Most Adaptive	Least Adaptive		Most Adaptive	Least Adaptive		Most Adaptive	Least Adaptive
Drugs	6	6	Drugs	11	7	Drugs	11	10
No Drugs	14	18	No Drugs	20	12	No Drugs	18	14
	$\chi^2 = .30$			$\chi^2 = .03$			$\chi^2 = .15$	

interpersonal situations change for better or worse while avoiding such situations results in stabilization at a marginal adjustment level.

The correlations among follow-up items are, in general, positive and significant. Yet it seems quite clear that the correlations are not of sufficient magnitude to warrant the assumption of a single criterion. Thus, the outcome of any treatment program is a function of the follow-up criterion chosen, which is based upon a value judgment. For instance, one would arrive at different conclusions concerning the efficacy of treatment programs in the current study if one arbitrarily decided that the criterion of hospitalization was more important than the criterion of employment.

But further complications in the criteria problem are revealed in the correlation matrix shown in Table 28. Within-hospital measures, with the exception of the group therapy rating, are not significantly related to follow-up. Thus, ward behavior, long an accepted criterion in treatment evaluation, is not significantly related to the three follow-up items chosen and performance in the hospital work situation is not significantly related to any other measure. Tests form a cluster of interrelated scales which are, by and large, not significantly related to other criteria. The relationships between follow-up criteria and the Group Therapy Rating Scale suggests that observations by trained personnel of intensive interpersonal situations within the hospital may be the best predictors of follow-up adjustment.

The fact that there are two relatively independent criteria clusters, (*a*) subjective self-evaluation as revealed in tests and (*b*) objective interpersonal evaluation as revealed in follow-up items and group therapy rating, caution against generalizing from test measures to within-hospital behavior or follow-up behavior. Implications of these relationships for treatment evaluation studies using a single criterion are obvious.

SUMMARY

This investigation attempted to evaluate the relative efficiency of four psychotherapeutic programs with three patient groups on a number of different criteria. All patients in the study lived on the same experimental ward. The four treatments were: (*a*) group living, in which all activities were group oriented, i.e., work, living arrangements, psychotherapy, etc., (*b*) group therapy and an individual work assignment, (*c*) individual therapy and an individual work assignment, and (*d*) a control group with an individual work assignment and no therapy. The three patient groups were: (*a*) a nonpsychotic group, (*b*) a short-term psychotic group, and (*c*) a long-term psychotic group. The design was a three by four factorial type analysis of variance with an N of eight in each cell and a total N of 96. All patients were tested and rated upon entry into the program and upon departure. The instruments were: (*a*) Ward Behavior, (*b*) MMPI, (*c*) TAT, (*d*) Holland Vocational Preference Inventory, (*e*) Q sort, and (*f*) a six-month follow-up questionnaire. In addition, all patients in group therapy were rated every 30 days on Finney's (1954) Therapy Scale and 70 of the 96 were rated on job performance, both upon entry and termination of the program.

The following results obtained:

1. The number of days in treatment was significantly different for both treatment and diagnostic groups. The individual therapy group took the greatest time, group living next, with group therapy and control the least and almost equivalent times. Diagnostically, long-term psychotics were in treatment longest, with short-term psychotics and nonpsychotics requiring the least and almost equivalent times.

2. Change score comparisons (initial–terminal test scores) were computed for variances and medians. Generally, significantly different variances occurred on most instruments between treatments and diagnoses. The control group usually had the smallest variance and the three psychotherapy groups the largest. Of the four treatments, group living or individual therapy usually showed the greatest variance. Variances were usually significantly

different between diagnoses with the long-term psychotics having the largest variances, followed in descending order by the short-term and nonpsychotic patients.

3. Tests of significance between change score medians revealed only the *Ma* scale of the MMPI yielding a significant difference between treatments. However, nine interactions were significant, and six of these followed the same pattern: nonpsychotics and short-term psychotics in the three groups receiving psychotherapy showed moderate to large adaptive change while long-term psychotics in the same three groups showed moderate to large change in a maladaptive scale direction. On the other hand, nonpsychotics and short-term psychotics in the control group showed relatively little change, and long-term psychotics in the same program showed moderate to large adaptive scale score change. Four scales showed significantly different median change between the diagnostic groups. On all of these scales the nonpsychotic and short-term psychotic groups changed more adaptively than the long-term psychotics.

4. Correlations between the nine follow-up items revealed three that were so highly intercorrelated as to be considered a single criterion. Other items were, in general, significantly correlated but not of sufficient magnitude to warrant criteria substitution. Tests of significance between treatments revealed significant differences in employment, with the three groups which had received psychotherapy having a higher percent of full-time employed than the control. The other eight follow-up items revealed no significant differences between treatments. When compared by diagnoses, six of the nine follow-up items significantly differentiated the groups. Overall, the short-term psychotics made the most adequate posthospital adjustments, nonpsychotics next, and the long-term psychotics the poorest.

5. Thirteen frequently used scales were selected for correlational analysis. They were: three follow-up items, (Rehospitalization, Employment, Degree of Illness), three within-hospital behavior measures, (Ward Behavior, Job Performance, and Group Therapy Behavior), TAT, Q sort, and *Sc, Es, A, R,* and *SIM* from the MMPI. Two relatively independent clusters of interrelated measures appeared: (*a*) three follow-up items and group therapy behavior and (*b*) MMPI scales of *Sc, Es, A, SIM,* Q sort, and TAT as a marginal member.

6. Other comparisons revealed that "global" follow-up adjustment was unrelated to time in treatment or receipt of tranquilizing drugs during treatment.

The results were discussed in relation to treatment programs, psychotherapy criteria, and further research.

REFERENCES

BARRON, F. An ego-strength scale which predicts response to psychotherapy. *J. consult. Psychol.*, 1953, **17**, 327–333.

EDWARDS, A. L. *Experimental design in psychological research.* New York: Rinehart, 1950. Pp. 195–197.

FAIRWEATHER, G. W., MORAN, L. J., & MORTON, R. B. Efficiency of attitudes, fantasies, and life history data in predicting observed behavior. *J. consult. Psychol.*, 1956, **20**, 58.

FERGUSON, J., MCREYNOLDS, P., & BALLACHEY, E. *The hospital adjustment scale.* Stanford, Calif.: Stanford Univer. Press, 1953.

FINNEY, B. C. A scale to measure inter-personal relationships in group psychotherapy. *Group Psychother.*, 1954, **7**, 52–66.

HOLLAND, J. L. *Manual for the HVPI.* Palo Alto, Calif.: Consulting Psychologists Press, 1958.

SHEER, D. E. A self-concept test: Reliability and validity data. Personal communication, 1955.

SHEFFIELD, F. D. Comment on a distribution-free factorial-design analysis. *Psychol. Bull.*, 1957, **54**, 426–428.

SIEGEL, S. *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill, 1956. Pp. 179–184.

SIMON, R. An investigation of the relationship between several measures of self-insight and adjustment, intelligence and personality characteristics. Unpublished doctoral dissertation, Syracuse Univer., 1952.

WELSH, G. S. Factor dimensions *A* and *R*. In G. S. Welsh & W. G. Dahlstrom (Eds.), *Basic readings on the MMPI in psychology and medicine.* Minneapolis, Minn.: Univer. Minnesota Press, 1956. Pp. 264–281.

WILCOXON, F. *Some rapid approximate statistical procedures.* New York: American Cyanamid, 1949. P. 5.

(Received March 24, 1959)

Vol. 74, No. 6 Whole No. 493, 1960

Psychological Monographs: General and Applied

EXPERIMENTAL STUDIES IN THE TRAINING OF ORIGINALITY[1]

IRVING MALTZMAN, SEYMORE SIMON, DAVID RASKIN, AND LEONARD LICHT[2]

University of California, Los Angeles

A BASIC difficulty in attempting to facilitate original thinking is that it may not occur at all or at such infrequent intervals that reinforcements cannot be administered with sufficient frequency to effect an increase in such behavior. Thus, a fundamental problem in the training of originality is to devise methods for increasing its occurrence in the first place, thereby permitting the operation of reinforcement. We are assuming that originality can be learned and that the same principles of conditioning hold as in other forms of operant behavior. Some of the problems attendant upon this assumption have been discussed elsewhere (Maltzman, 1960). For our present purposes we need indicate only that by originality, or original thinking, we mean behavior that occurs relatively infrequently, is uncommon under given conditions, and is relevant to those conditions. In order, then, to facilitate the occurrence of original behavior, techniques must be devised for evoking many uncommon responses. Such training may then produce a disposition to give uncommon responses in other situations.

One procedure, used by Maltzman, Bogartz, and Breger (1958) is to repeatedly evoke different associations to the same stimulus words in a free association situation. This procedure prompts the subject to emit responses relatively low in the response hierarchy elicitable by each stimulus. They found that subjects receiving this training with intermittent reinforcement of uncommon responses were significantly more original on a new list of words than subjects without this training. A second experimental group which did not receive verbal reinforcement approached significance in comparison with the control group. When instructions to be original were added to the experimental and control treatments they showed a marked increase in the originality of their associations, and both experimental groups were now significantly more original than the control group. The Unusual Uses Test of originality devised by Guilford and his associates (1950) was administered following the free association test. Equivocal evidence for transfer of the association training effects to this situation was obtained in the form of a significant triple order interaction.

The essential features of the procedure responsible for the facilitation of original responses to new stimuli, however, cannot be determined from that study. The purpose of the present series of experiments, in part, is to determine what those relevant features are.

EXPERIMENT I

It may be that the important characteristic of the procedure employed by Maltzman, Bogartz, and Breger (1958) is not evocation of responses relatively low in the hierarchy of each stimulus word, as was assumed, but simply the production of many different responses. The purpose of this first experiment, therefore, was to determine whether facilitation of uncommon responses to new stimuli would occur when different responses were evoked by the presentation of different stimuli rather than by repetition of the same stimuli. Two different lists of stimulus words were employed

---

[1] This research was supported by Contract Nonr 233(50) between the Office of Naval Research and the University of California. Reproduction in whole or in part is permitted for any purpose of the United States Government.

[2] We are indebted to William Bogartz, Nira Kozak, and Bonnie Simon for their assistance in various stages of this research.

for this purpose. One list consisted of relatively common and the other relatively uncommon words, in order to determine whether the frequency of usage of the words would have a differential effect upon the uncommonness of test responses. Another control group not employed in the previous study (Maltzman, Bogartz, & Breger, 1958) was added to this experiment. This group responded in the same fashion to repeated presentations of a given stimulus word. Such a condition would enable us to determine whether the number of responses per se, or the number of different responses, is the relevant variable influencing test performance.

Method

Subjects. The subjects (*S*s) were 292 students drawn from introductory psychology classes.

Stimulus materials. The stimulus words used in the initial list and the test list for all groups were selected from the norms obtained by Wilson (1942), and are the same as those used in a previous study (Maltzman, Bogartz, & Breger, 1958). Evocation of a relatively small number of different responses in a free association situation was the criterion for the selection of these stimulus words. The different training words presented to two of the groups following the initial list were selected from the Thorndike-Lorge count (1944). The list of relatively common words were chosen from among the 500 most frequently occurring words in the count. The list of relatively uncommon words were selected from words occurring not more than 6 times per 1,000,000. Guilford's Unusual Uses Test (1950) was administered after the completion of the free association test list.

Procedure. The *S*s were treated in identical fashion on the initial presentation of the free association training list, the first 25 stimulus words for all the groups. At the start of the experiment the *S*s received the usual free association instructions to respond as quickly as possible with the first word that came to mind.

After completion of the initial 25-word training list one control group (C) was given the test list of 25 new stimulus words. The previous instructions were repeated before presentation of the test list, as they were for all the other groups. A second control group (C_R) received five additional presentations of the same training list with instructions to try to give the same response to a given stimulus word each time. They received the test list following the last repetition of the initial training list.

Two experimental groups received a single presentation of the initial 25-word training list followed immediately by 125 different words. One group (X_L) received a list of words with a low frequency count, while the other group (X_H) received words with a high count. They were given the test list following the completion of the entire training list.

After completion of the training list a third experimental group (X) received five additional presentations of the list with instructions to respond as quickly as possible to each word, but to give a different response than the one used before. The test list was administered following the last repetition of the training list.

All groups received the Unusual Uses Test immediately after the completion of the free association test list.

The experiment was conducted in group form. Stimulus words for the free association training and test were read aloud by *E*, and *S* recorded his responses in a booklet provided. Each page of the booklet had 25 numbered blank spaces for his responses. A stimulus word was presented every 5 sec. during the initial training list and the test list for all groups. The additional training words presented groups X_H and X_L were presented at the same rate. Repetitions of the initial training list words for groups X and C_R were presented at 10 sec. intervals.

Results

The frequency with which the different responses occurred to each stimulus word was determined for the free association training and test lists. *S*s' responses were then scored on the basis of the obtained frequencies, and each *S* was assigned an originality score which was the mean frequency of his responses. A low score therefore represents high originality. A few *S*s failed to respond to a stimulus word or recorded an illegible response, but no *S* gave less than 24 readable responses. Their score was based on the number of legible responses recorded.

Table 1 shows the mean originality scores obtained by the various groups on the first presentation of the training list, the pretest. A simple analysis of variance showed that they did not differ significantly prior to the introduction of the different treatments ($F = 2.22$; $df = 4{,}287$; $P > .05$). Homogeneity of variance was indicated by the Hartley test.

Table 2 shows the mean originality scores obtained by the various groups on the free association test list. An analysis of covariance of the free association test scores was

TABLE 1

Mean Pretest Free Association Originality Scores in Exp. I

Cond.	*N*	Mean	*SD*
C	64	91.69	19.74
$\mathrm{C_R}$	62	92.70	16.42
X	60	86.39	25.89
$\mathrm{X_H}$	60	97.06	19.35
$\mathrm{X_L}$	46	95.89	20.64

planned in order to increase the precision of the experiment. However, a test of the homogeneity of regression of the test on the training scores indicated that this assumption of the analysis could not be met ($P < .005$).

A simple analysis of variance of performance on the free association originality test was therefore conducted. An F of 14.95 was obtained ($df = 4{,}287$; $P < .001$). The results of t tests showed that the two control groups did not differ significantly from each other and the three experimental groups did not differ significantly. However, each of the three experimental groups was significantly more original than each of the control groups, $P < .001$.

Originality scores for performance on the Unusual Uses Test were obtained for each *S* by a procedure somewhat similar to the one employed for the free association test. The frequency with which different uses occurred for each of the six different objects mentioned in the test was first determined. In the previous study by Maltzman, Bogartz, and Breger (1958) an originality score was obtained for each *S* by counting the number of responses he gave to each test item that occurred less than 10% of the time in the obtained sample. This procedure was not feasible in the present study which employed more than twice the number of *S*s, since unique responses typically comprised more than 10% of the responses to each item. The originality score for each *S* therefore was taken as his total number of unique responses. Any response was defined as unique if it occurred only once in the sample to a given test item. A second score obtained for each *S* was his total number of nonunique or common responses. This may be taken as a measure of fluency.

Table 3 shows the mean number of unique responses obtained by the various groups on the Unusual Uses Test. The results from one *S* in Group $\mathrm{C_R}$ were discarded, because he incorrectly recorded his uses in the free association test booklet rather than the booklet provided for the Unusual Uses Test. Since Hartley's test indicated that the variances of these scores were heterogeneous ($P < .01$) and the distributions were highly skewed, it was decided to adopt the .01 level for rejection of the null hypothesis. An analysis of covariance was not conducted because of the absence of a significant correlation between unique responses and originality on the free association pretest.

A simple analysis of variance of the originality measure obtained from the Unusual Uses Test gave an F of 18.74 ($df = 4, 286$; $P < .001$). Since the variances were significantly different, half the appropriate degrees of freedom were used in the t tests of the differences between groups. Results of these tests showed that Group X gave significantly more unique responses

TABLE 2

Mean Originality Scores on Free Association Test in Exp. I

Cond.	Mean	*SD*
C	76.38	23.18
$\mathrm{C_R}$	70.90	20.99
X	54.49	18.42
$\mathrm{X_H}$	53.40	24.57
$\mathrm{X_L}$	56.44	18.64

TABLE 3

Mean Number of Unique Uses in Exp. I

Cond.	Mean	*SD*
C	2.58	2.30
$\mathrm{C_R}$	1.03	1.06
X	4.75	3.54
$\mathrm{X_H}$	2.27	2.08
$\mathrm{X_L}$	2.72	2.36

TABLE 4

MEAN NUMBER OF COMMON USES IN EXP. I

Cond.	Mean	*SD*
C	14.44	4.84
C_R	14.59	4.58
X	15.65	5.09
X_H	15.42	4.18
X_L	15.62	4.34

than any of the other groups, $P < .001$. Groups X_H, X_L, and C did not differ significantly. However, Groups C and X_L were significantly more original than Group C_R at the .001 level. Group X_H was significantly more original that C_R at the .01 level. A rank order analysis of variance yielded comparable results ($H = 52.00$; $df = 4$; $P < .001$).

Table 4 shows the mean number of non-unique or common responses obtained by the different groups on the Unusual Uses Test. Hartley's test indicated that the variances were homogeneous. A simple analysis of variance yielded an *F* of .92, indicating that the groups did not differ significantly in terms of their fluency of common responses on the Unusual Uses Test. Analysis of covariance was again contraindicated by the absence of a significant correlation between frequency of common uses and originality on the free association pretest.

A basic assumption in the use of the training procedure employed with Group X is that it would increase the uncommonness of the responses to the training words. Fig. 1 shows the mean percentage of unique responses occurring on successive repetitions of the training list. It was necessary to employ the percentage of unique responses rather than the mean or median weighted frequency scores, because of the highly skewed distributions of scores obtained for the 25 training words. By the third repetition over half the responses of the average *S* were unique. It is clear from Fig. 1 that the training method employed with Group X had the desired effect of inducing an increase in the uncommonness of the responses emitted.

Group C and Group X in the present experiment are similar to conditions employed in the study by Maltzman, Bogartz, and Breger (1958). In the latter experiment the difference in originality between these two groups only approached significance on the free association test, but was significant when combined with verbal reinforcement or instructions to be original. In the present experiment Group X showed a highly significant increase in free association originality in the absence of verbal reinforcement or differential instructions to be original. In the previous experiment only equivocal evidence of facilitation on the Unusual Uses Test was obtained. Here, the training administered Group X in the free association situation produced a highly significant facilitation of originality on the Unusual Uses Test.

The different results obtained in the two experiments may only be due to the increase in the size of the sample in the present experiment. However, it is possible that the change from an individual to a group experiment also contributed to the difference in results. Conducting a free association experiment in group form grants a degree of anonymity to *S*'s responses which is not the case when he is run on an individual

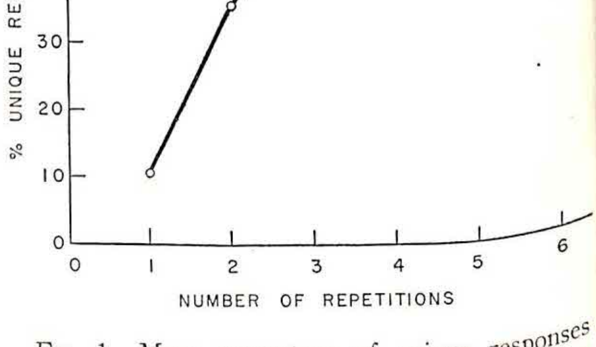


FIG. 1. Mean percentage of unique responses obtained by Group X with free association training in Exp. I.

basis. In the group experiment there also are necessarily more different cues present that are potentially the occasion for many responses less likely to occur when *S* is alone with *E* in a bare room. Both of these characteristics may have contributed to the appearance of a training effect in this experiment of a greater magnitude than the one reported in the earlier study (Maltzman, Bogartz, & Breger, 1958). In any case the magnitude of the training effect obtained in this study is consistent with the results of the subsequent experiments reported here.

Groups X_H and X_L were employed in order to determine whether evoking many different responses would produce the same effect on test performance as the training provided Group X. The assumption was that the occurrence of any given response increases the disposition of responses associated with it to occur. Evidence of such an effect has been reported by Judson and Cofer (1956) and by Judson, Cofer, and Gelfand (1956). Providing for the occurrence of many different responses during free association training should therefore increase the reaction potential of many more different responses on the test list as compared with the control groups. That an effect of this sort may be operating is suggested by the results of the free association test where the three experimental groups do not differ significantly, but are significantly superior to the control groups.

However, the results of the Unusual Uses Test in terms of originality scores indicates that the training administered Group X involves more than the arousal of many intraverbal associations. On this test, Group X gave significantly more unique responses than Groups X_H and X_L as well as the two control groups. These results indicate that the procedure of repeated evocation of different responses to the same stimulus words produces a more general disposition to produce uncommon responses than the procedure of evoking different responses to different stimuli. The effect of this training, furthermore, was restricted to increasing the number of uncommon uses. There were no significant differences among the groups in terms of their fluency of common responses.

The significantly poorer performance on the free association originality test of Group C_R than the three experimental groups indicates that the increased originality of the experimental groups on this test cannot be attributed to the number of responses evoked per se or to habituation factors. The three experimental groups and Group C_R emitted the same number of responses during training. The difference obtained must be due to the uncommonness of responses per se emitted by the experimental groups or the uncommonness of responses emitted to the given stimulus words.

The fact that Group C_R gave significantly fewer unique uses than Group C indicates that the training in the free association situation of repeatedly evoking the same response to a given stimulus interferes with the tendency to give uncommon uses. To some extent, the repeated occurrence of common responses to the same stimuli increases the tendency to emit common responses in other situations, just as the repeated occurrence of uncommon responses to the same stimuli increases the disposition for original behavior in other situations.

EXPERIMENT II

The purpose of this next experiment was to further explore possible relevant variables and alternative methods of facilitating originality. Two groups of *S*s were exposed to uncommon associations which they simply read, in order to determine whether emission of uncommon responses by *S* is a necessary condition of the originality training procedure. In an attempt to explore the extent to which the method of facilitating originality is peculiar to the training materials employed, a group of *S*s was given repeated presentations of items from the Unusual Uses Test and tested with the free association materials. Conditions X and C of the previous experiment, which we shall hereafter call the standard experimental training and control conditions, were also employed in this experiment.

Method

Subjects. The *S*s were 251 students drawn from introductory psychology classes.

Stimulus materials. The stimulus words used in the initial training and subsequent test lists are the same as those in Experiment I. However, their function was reversed in the present experiment. The list previously employed as the training list was used as the test list in this experiment, and the test list in the previous experiment was used as the training list. Guilford's Unusual Uses Test was administered after the completion of the free association test list to all groups except X_2.

Procedure. All *S*s were presented the initial training list and the final test list of stimulus words in the same manner, including the usual free association instructions to respond as quickly as possible with the first word that came to mind.

The standard control group (C) received only a single presentation of the training list followed by the test list of words. The standard experimental condition (X_1), as in the previous study, received five additional presentations of the training list with instructions to respond as quickly as possible, but to give a different response than the one used before to each stimulus.

A second experimental group (X_2) was also asked to respond differently to repeated presentations. However, the stimuli employed were the six items from the Unusual Uses Test. The *S*s were instructed to write a different use following the oral presentations of each item. Five repetitions of the items were administered. A third experimental group (X_3) was presented with a booklet of 125 pairs of items. The *S*s in this condition were instructed to underline the member of each pair which goes more readily with the stimulus word orally presented to them. The stimulus words were five repetitions of the training list. Response pairs were unique responses to these words culled from the results of Exp. I.

The final experimental group, X_4, received the same response pairs as Group X_3, but without the preceding stimulus words. These *S*s were instructed to underline the member of each pair that they thought to be more familiar.

Stimulus words for the free association training and test lists were read aloud by *E*, and *S* recorded his responses in a booklet provided by *E*. The Unusual Uses Test was presented in written form to all groups except Group X_2. Stimulus words were presented every 5 sec. during the initial free association training and test. Repeated presentations of the stimuli for originality training occurred at 10 sec. intervals. Fifteen sec. were permitted for the response to each unusual uses training item in Group X_2.

Results

Frequency distributions of the responses to each of the training words in Exp. I were employed as norms for scoring the responses to these words which were employed as the test list in the present study.

TABLE 5

MEAN PRETEST FREE ASSOCIATION ORIGINALITY SCORES IN EXP. II

Cond.	*N*	Mean	*SD*
X_1	64	87.72	20.02
X_2	46	83.56	21.37
X_3	48	87.85	18.36
X_4	39	85.66	18.98
C	54	83.48	23.18

The frequency with which the different responses occurred to each of the training words in the present experiment was determined, and *S*s' responses were scored on the basis of these frequencies.

Table 5 shows the mean originality scores obtained by the various groups on the first presentation of the training list. A simple analysis of variance indicated that the groups did not differ significantly prior to the introduction of the experimental treatments ($F = .55$). Homogeneity of variance was indicated by the Hartley test.

Table 6 shows the mean originality scores obtained by the various groups on the free association test. Appropriate tests indicated that the variances of the test scores were homogeneous as well as the regression of test scores on training scores. The regression of the test scores on the initial training scores was .69.

A simple analysis of covariance was therefore conducted in which the originality test scores were adjusted for initial differences on the training list. An F of 31.86 was obtained ($df = 4, 245; P < .001$). The results of t tests indicated that Group X_1

TABLE 6

UNADJUSTED AND ADJUSTED (MEAN′) FREE ASSOCIATION TEST ORIGINALITY SCORES IN EXP. II

Cond.	Mean	*SD*	Mean′
X_1	55.09	20.44	53.73
X_2	80.02	25.88	81.54
X_3	87.82	19.58	86.38
X_4	84.37	21.90	84.43
C	80.67	24.08	82.23

TABLE 7

MEAN NUMBER OF TRANSFORMED ($\sqrt{X + .5}$) UNIQUE USES IN EXP. II

Cond.	Mean	*SD*
X_1	1.65	.56
X_3	1.35	.58
X_4	1.23	.54
C	1.17	.43

differed significantly from each of the other groups ($P < .001$), whereas these four groups did not differ significantly from each other.

The frequency with which different uses were offered for each item in the Unusual Uses Test was determined from the norms of Exp. I, and an originality score assigned each *S* based on the number of unique responses he submitted. A second score obtained for each *S* was his total number of nonunique or common responses. Since the score distributions for unique responses were skewed, and the variances were heterogeneous, as in the previous study, the transformation $\sqrt{X + .5}$ was applied to them.

Table 7 shows the mean number of transformed unique responses obtained by the various groups receiving the Unusual Uses Test. The variances of the transformed scores were homogeneous as well as the regression of unique responses on the free association pretraining scores. An analysis of covariance of the unique responses regressed on the pretraining scores yielded an *F* of 10.31 ($df = 3, 207$; $P < .001$). However, since the regression was only .02 no increase in precision over a simple analysis of variance was obtained, and the adjusted means which differ only slightly from the unadjusted means will not be presented.[3] Individual comparisons among the groups by means of *t* tests showed that the standard experimental training condition, Group X_1, was significantly more original than each of the other conditions ($P < .001$). The other conditions did not differ significantly from each other.

Since the distributions of scores were highly skewed due to the large number of failures to give unique responses, the transformation could not normalize the data. A rank order analysis of variance was therefore conducted. Results comparable to the previous analysis were obtained. The *H* was 29.82 ($df = 3$; $P < .001$).

Table 8 shows the mean number of common uses obtained by the various groups on the Unusual Uses Test. The variances were homogeneous as well as the regression of common responses on the free association pretraining scores. An analysis of covariance of the common responses regressed on the pretraining scores yielded an *F* of 5.41 ($df = 3, 207$; $P < .001$). However, since the regression was only $-.02$ no increase in precision over a simple analysis of variance was obtained. Individual comparisons among the conditions by means of *t* tests showed that the three experimental groups did not differ significantly from each other. However, each of these three groups gave significantly more common responses than the standard control group ($P < .05$).

Results obtained in this experiment for Condition X_1 confirm the results of Exp. I that repeated evocation of different responses to free association stimuli is followed by an increase in the originality of responses to new stimulus words and to items on an Unusual Uses Test as compared to a control group (C) in which such train-

TABLE 8

MEAN NUMBER OF COMMON USES IN EXP. II

Cond.	Mean	*SD*
X_1	20.09	5.17
X_3	19.26	5.91
X_4	21.08	5.63
C	17.03	5.18

---

[3] The statistical analyses of the results from this experiment were performed on an IBM 650 electronic computer. The results of the analysis of covariance are reported even though they are essentially the same as from an analysis of variance, since the computer program was for covariance, and the computations of the regressions were obtained at the same time.

ing is omitted. Conditions X_3 and X_4 were introduced in order to determine whether the uncommon responses must be evoked by the stimulus words in order for the training to be effective. In the former condition uncommon responses were evoked as textual responses (Skinner, 1957) following the stimulus words, whereas in Condition X_4 only the textual responses occurred. Neither condition was followed by significant increases in originality on the tests employed as compared with Group C, and they obtained significantly lower scores than the standard experimental condition, Group X_1. These results show that with the stimulus materials employed, uncommon responses must be evoked as intraverbal responses by the stimuli if there is to be an increase in originality in the test situations.

It may be argued perhaps that the associations provided Groups X_3 and X_4 are not uncommon. Although this is a possibility it seems rather unlikely, since each association occurred only once in a similar sample of approximately the same number of *S*s. Furthermore, since a pair of uncommon responses was given for each stimulus word or ordinal position, twice as many uncommon responses were available to Groups X_3 and X_4 than to X_1. Thus, even though some of these associations may have been common to the former groups they still should have received at least as many uncommon associations as *S*s in Group X_1.

Another hypothesis which may be suggested to account for the difference among the experimental conditions is that the procedure employed with Group X_1 induces the inhibition of common responses as the result of repeated presentations of the same stimulus words and the accompanying instructions, thereby permitting the appearance of uncommon responses. Thus, the facilitation of uncommon responses in the test situations occurs at least in part because of a disposition to inhibit common responses as well as an increase in the excitatory potential of uncommon responses. Conditions X_3 and X_4 do not provide for the inhibition of common responses, but evoke the uncommon responses directly as textual responses.

Evidence apparently contradicting this interpretation stems from another study (Maltzman, Brooks, Bogartz, & Summers, 1958) which found that uncommon uses occurring as textual responses may facilitate performance on the Maier two-string problem. Examination of the latter experiment suggests, however, that the results obtained there are not necessarily in conflict with the inhibition interpretation. In the latter experiment the uncommon textual responses were evoked in the presence of stimulus objects subsequently found in the two-string problem. The excitatory potential of uncommon responses to the test stimuli could therefore be increased by the training procedure employed. But in the present experiments the training and test stimuli were different. This suggests that a training method evoking uncommon intraverbal responses will produce a more general effect than training with uncommon textual responses, and implies further that if uncommon uses for objects not found in the two-string problem were to be employed, facilitation of problem solutions would not occur.

Although Conditions X_3 and X_4 did not show a significant increase in originality in the two test situations, they did yield a significant increase in common responses on the Unusual Uses Test, an effect also obtained in Condition X_1. This finding suggests that uncommon responses, whether textual or intraverbal, tend to facilitate the occurrence of other responses, to facilitate fluency. However, since this result was not anticipated, the experiment was not designed with an adequate control group to determine whether an increase in common uses is peculiar to the evocation of uncommon responses during training or whether such an increase would occur as well if common responses were initially evoked. It should be noted that Condition X_1 in Exp. I did not show a significant increase in common responses as compared to the control group, while highly reliable increases in original responses were obtained. This discrepancy in the results of the two experiments suggests that the present findings of increases in common responses should be accepted with caution. The results of subsequent experiments are equally inconsistent.

EXPERIMENT III

The purpose of this experiment was to determine whether Group X_2 in Exp. II failed to show a significant training effect due to the relatively small amount of training it received, five repetitions of the six items from the Unusual Uses Test. In contrast, the standard experimental condition received 25 stimulus items repeated five times. Before it can be concluded that unusual uses items cannot be used as training materials for the facilitation of originality, the effects of more extensive training with these materials must be explored. However, amount of training can be varied in two ways, by increasing the number of items, and by increasing the number of repetitions of the items. Both variations were employed in this experiment.

Method

Subjects. The *S*s were 162 students drawn from introductory psychology classes.

Stimulus materials. The same initial and final free association lists were employed as in Exp. II. Items from the Unusual Uses Test previously employed were also used in this experiment. Six additional items were used in two of the conditions. These items were, a brick, a pair of scissors, a pencil, a five-cent piece, a light bulb, and a bar of soap.

Procedure. Four different groups of *S*s were employed similar to Group X_2 in Exp. II. They all received a single presentation of the free association training and test lists. Group A received 10 repetitions of the six items while Group C_1 received 20 repetitions of these items. Group B received five repetitions of 12 items and Group D received 10 repetitions of these 12 items prior to presentation of the free association test.

TABLE 9

MEAN PRETEST FREE ASSOCIATION ORIGINALITY SCORES IN EXP. III

Cond.	*N*	Mean	*SD*
A (6 × 10)	43	70.62	30.43
C_1(6 × 20)	38	82.95	21.27
B (12 × 5)	43	78.75	24.10
D (12 × 10)	38	84.71	23.09

TABLE 10

MEAN FREE ASSOCIATION TEST ORIGINALITY SCORES IN EXP. III

Cond.	Mean	*SD*
A (6 × 10)	65.39	28.27
B (12 × 5)	68.27	21.23
C_1(6 × 20)	79.17	24.20
D (12 × 10)	77.22	29.32

Results

Performance on the free association training and test lists was analyzed and scored in the same fashion as in the previous experiments. Table 9 shows the mean originality scores obtained by the four groups on the training list. A two-way analysis of variance showed that the column means, 60 vs. 120 responses, differed significantly prior to the introduction of the experimental treatments ($F = 5.21$, $df = 1, 158$; $P < .05$). Factors responsible for this initial bias are not at all apparent. Table 10 shows the unadjusted mean originality scores obtained by the four conditions of this experiment on the free association test list. Appropriate tests showed that the variances of the training and test means were homogeneous. A test of homogeneity of regression of the test scores on training scores, however, was significant ($F = 4.76$; $df = 3, 120$; $P = .01$), thus precluding an analysis of covariance.

Fortunately, however, the assumptions of analysis of covariance were met in the comparison of primary interest.

Comparison of the two conditions in this experiment given six items with 10 and 20 repetitions, conditions A and C_1, respectively, with conditions X_2 and C of the previous study would indicate whether the number of repetitions has a significant facilitating effect upon free association originality. Table 11 presents the adjusted free association test means of these four groups. Their pretraining mean originality scores do not differ significantly. Their

TABLE 11

UNADJUSTED AND ADJUSTED (MEAN') ORIGINALITY TEST SCORES FOLLOWING DIFFERENT AMOUNTS OF TRAINING IN EXPS. II AND III

Cond.	Mean	Mean'
C	80.66	78.81
$X_2(6 \times 5)$	84.36	81.07
A (6×10)	65.39	72.03
$C_1(6 \times 20)$	79.17	77.67

variances and regression on pretraining scores likewise were homogeneous. The regression was .66. A simple analysis of covariance gave an *F* of 1.62 ($df = 3, 169$; $P > .05$). Thus, no evidence of a facilitating effect of training with unusual uses was obtained.

Results obtained with Group X_2 of Exp. II and the four groups of Exp. III indicate that the method of facilitating original responses by repeated evocation of different responses to the same stimuli does not work with all materials.

A basic assumption of the approach employed here is that if original behavior can be induced to occur repeatedly through some training procedure, it will be more likely to occur in new situations. The training method employed, therefore, must evoke many uncommon responses. If this condition is not met, transfer of originality to new stimulus situations cannot occur. It has previously been determined that repeated evocation of different responses to free association materials does lead to increased originality on successive training trials.

The mean percentage of unique uses responses occurring on successive repetitions of the training list was therefore determined for Group A. Fig. 2 shows the results of analyzing the training data, and suggests that a reliable increase in unique responses may have occurred during training. A trend analysis was therefore conducted. It yielded a significant trials effect ($F = 2.60$; $df = 9, 378$; $P = .01$) and *S*s' effect ($F = 4.24$; $df = 42, 378$; $P < .001$). Thus, a significant increase in the percentage of unique uses occurred with successive repetitions of the training list. However, a comparison of the effects of training with free association stimulus words and uses, Figs. 1 and 2, indicates that the former produced considerably more unique responses. The mean percentage of unique responses is higher in the former case after a single repetition of the list of 25 words than the latter after 10 repetitions of the 6-item list. This is probably due to the relatively low percentage of unique uses on the first presentation of the unusual uses list. A low operant level would retard the growth of a training effect.

Another possibility is that the kinds of responses evoked by the stimulus items is of critical importance. The free association stimulus materials evoke different discrete responses on each presentation of a stimulus item. In contrast, entire phrases are evoked by the unusual uses items, many of the components of which are common to a large number of the phrases. This would have the effect of reducing the uncommonness of the phrases evoked during training, since many of the component verbal responses occur repeatedly.

In addition, it appears that a relatively limited number of different uses are available for each item. Given the instructions to respond to each repetition of an item and the inability of *S*s to do so with a different operant, they respond with essentially synonymous uses repeatedly. It is apparent that considerably more experimental re-

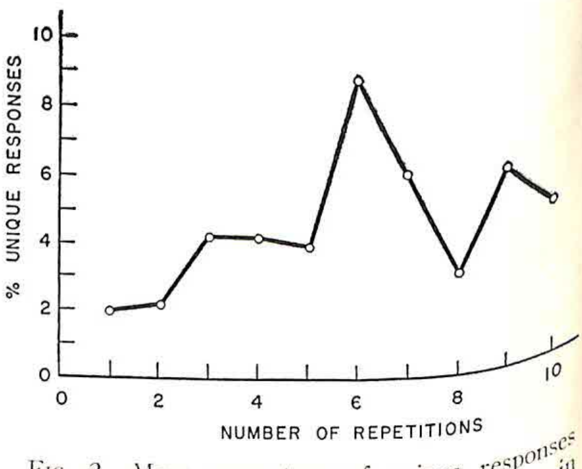


FIG. 2. Mean percentage of unique responses obtained by Group A with unusual uses training in Exp. III.

search is needed in order to adequately determine the training conditions and characteristics of stimulus materials that are necessary to facilitate originality in test situations.

EXPERIMENT IV

Condition X_2 in Exp. II indicated that originality training with unusual uses items does not affect performance on the free association test of originality. That is, the experimental procedure previously found to be effective, the evocation of different responses to the same free association stimuli materials, did not have a comparable effect when the stimuli were the names of common objects, and the responses were unusual uses for these objects. Another study, Exp. III, which varied the number of items and the number of repetitions of the items, likewise failed to produce a significant change in the uncommonness of responses on the free association test of originality. It was concluded that repeated presentations of the unusual uses items was ineffective because it did not produce a large enough increase in the uncommonness of the responses evoked during training.

The failure to modify originality test performance by training with unusual uses leaves another important problem unanswered. If, as we have assumed, originality may be learned by the same principles as other forms of operant behavior, it should be a function of the amount of training administered. But since training with the unusual uses items failed to produce any significant effect on test performance, the latter hypothesis could not be adequately studied in the above experiments.

The purpose of the present experiment, therefore, was to determine whether performance on tests of originality would vary as a function of the number of repetitions of the training list. Training was conducted in the free association situation, since reliable effects upon originality have previously been obtained under these conditions. However, in order to increase the generality of the results of this series of studies somewhat, new training and test words were employed as free association stimuli.

Method

Subjects. The *S*s were 201 students drawn from introductory psychology classes.

Stimulus materials. The stimulus words used in the initial training and free association test lists were drawn principally from Woodworth and Wells (1911), and were selected on the basis of their tendency to evoke opposites. This would have the effect of producing a high communality of responses to each stimulus word, i.e., relatively few different responses would be evoked. Table 12 shows the free association training and test words employed. There were 24 words in the training

TABLE 12

STIMULUS WORDS EMPLOYED IN THE FREE ASSOCIATION TRAINING AND TEST LISTS IN EXP. IV

Training List		Test List
1. yes	13. dirty	1. summer
2. good	14. wrong	2. out
3. east	15. empty	3. above
4. day	16. after	4. north
5. love	17. deep	5. top
6. soft	18. large	6. wet
7. far	19. tall	7. up
8. smooth	20. open	8. long
9. early	21. come	9. big
10. dead	22. male	10. front
11. asleep	23. cold	
12. lost	24. low	

list and 10 in the test list. Guilford's (1950) Unusual Uses Test was administered after completion of the free association test list.

Procedure. The four principal groups were treated in identical fashion on the initial presentation of the free association list, the first 24 stimulus words for these groups. The usual free association instructions employed in earlier experiments of this series were administered. The standard control condition (C) received the test list of 10 new stimulus words following completion of the initial 24-word training list. Since the free association training and test words were selected because they gave a high proportion of opposite responses, it was considered possible that responding to the first 24 stimulus words might induce a tendency to emit opposites on the free association test list. A second control group (C_X) was therefore employed to determine whether this effect would occur. *S*s in this group received only the free association test list, omitting the 24-word training list.

Three different experimental groups received one, five, or ten repetitions of the training list prior to presentation of the free association and Unusual Uses tests. The latter test was employed in a slightly modified form from the standard procedure. The items were presented individually rather than in the usual booklet form with a time limit for each of the two sections of the test. Under the latter conditions *S*s tend to devote a disproportionate amount of time responding to the first item in each section, thereby giving relatively few responses to later items in the test. Presenting the items individually would tend to equate the number of responses given to each item.

Results

Since these training and test words had not been used in the previous experiments, response norms for the stimuli had to be constructed. The frequency with which the different responses occurred to each stimulus word in the training and test lists was therefore determined and *S*s' responses were then scored on the basis of the obtained frequencies.

Table 13 shows the mean originality scores obtained by the various groups on the first presentation of the training list. Homogeneity of variance was indicated by the Hartley test. But a simple analysis of variance indicated that the groups differed significantly prior to the introduction of the experimental treatments ($F = 4.81$; $df = 3, 155$; $P < .005$). The results of *t* tests indicated that Group X_5 was significantly superior to each of the other groups on the pretest ($P < .05$).

TABLE 13

MEAN PRETEST FREE ASSOCIATION ORIGINALITY SCORES IN EXP. IV

Cond.	*N*	Mean	*SD*
C	44	90.01	23.89
X_1	42	93.68	20.31
X_5	39	73.14	33.54
X_{10}	34	88.14	20.41

Fortunately, the regression of the free association test scores on the pretest scores was considerable, .87. Tests of heterogeneity of regression and homogeneity of variance were not significant, permitting an analysis of covariance in which the originality test scores were adjusted for the initial differences in the training list. Table 14 shows the mean unadjusted and adjusted originality scores obtained by the various groups on the free association test. A simple analysis of covariance was conducted and an *F* of 20.94 was obtained ($df = 3, 154$; $P < .001$). The results of *t* tests showed that Groups X_5 and X_{10} did not differ significantly. However, all other differences among the four groups were statistically significant ($P < .01$).

A *t* test was conducted between Groups C and C_X ($N = 42$; $M = 73.47$; $\sigma = 36.30$) in order to determine whether the emission of opposites on the training words would increase the tendency to give opposites on the test list. If such were the case, Group C would have significantly higher test scores than Group C_X, since opposite responses

TABLE 14

UNADJUSTED AND ADJUSTED (MEAN') FREE ASSOCIATION TEST ORIGINALITY SCORES IN EXP. IV

Cond.	Mean	*SD*	Mean'
C	82.88	28.91	79.60
X_1	68.76	33.95	62.29
X_5	36.03	28.12	47.43
X_{10}	46.03	33.62	44.43

were the most frequently occurring responses to the stimulus words. This was not the case ($t = 1.32$; $df = 84$; $P > .10$).

The frequency of unique responses on the Unusual Uses Test was determined for each *S* on the basis of the norms for this test derived from Exp. I. A second score obtained for each *S* was his total number of nonunique or common responses. Since the score distributions for unique responses are skewed, and the variances heterogeneous, the transformation $\sqrt{X + .5}$ was applied to them.

Table 15 shows the mean number of transformed unique responses obtained by the various groups receiving the Unusual Uses Test. Hartley's test indicated that the variances of the transformed scores were homogeneous. Since, as in previous studies, there was no significant correlation between the number of unique or common responses and originality on the free association pretest list, a simple analysis of variance rather than covariance was conducted. The results of the analysis were significant ($F = 5.82$; $df = 3, 155$; $P < .001$). The results of *t* tests indicated that the three experimental groups did not differ significantly from each other. Each of the three experimental groups gave significantly more unique responses than Group C ($P < .05$ for Group X_5 vs. C, and $P < .001$ for the other two experimental groups vs. Group C).

Table 16 shows the mean number of common responses obtained by the various groups on the Unusual Uses Test. The variances were homogeneous as indicated by the Hartley test. A simple analysis of variance was conducted, yielding a significant result ($F = 4.62$; $df = 3, 158$; $P < .01$). The results of *t* tests showed that Group X_5 gave significantly more common uses than each of the other two experimental groups ($P < .05$) as well as Group C ($P < .01$). No other differences among the groups were significant.

TABLE 15

MEAN NUMBER OF TRANSFORMED ($\sqrt{X + .5}$) UNIQUE USES IN EXP. IV

Cond.	Mean	*SD*
C	1.01	.45
X_1	1.43	.54
X_5	1.24	.45
X_{10}	1.43	.64

TABLE 16

MEAN NUMBER OF COMMON USES IN EXP. IV

Cond.	Mean	*SD*
C	19.48	5.42
X_1	21.19	5.06
X_5	23.85	6.13
X_{10}	20.94	4.57

Results from the free association test of originality indicate that each of the three experimental groups was more original than the control group. Our analysis, however, cannot indicate the precise nature of the trend produced by varying amounts of training, and the experiment was not designed for this purpose, but only to determine whether or not some form of trend exists. The data obtained show that reliable changes in free association originality may be produced by varying the amount of training in the standard experimental procedure. The particular form of the trend, however, is no doubt a function of the length of the training list as well as the number of its repetitions. Thus, a factorial experiment must be conducted in which both the length of the training list and the number of repetitions are varied before it may be possible to specify the function that obtains between amount of training and originality.

Although all three experimental groups gave significantly more unique responses than Group C, the experimental conditions did not vary as a function of the number of repetitions received. The basis for the absence of a trend among the experimental conditions on the Unusual Uses Test as contrasted with the free association test cannot be determined, and one can only speculate at this point as to why this discrepancy was obtained. A possibility may be the lack of power of the simple analysis of variance as compared with covariance and/or sampling error in the case of Group X_5. These same factors may also be responsible for the significantly greater number of common uses offered by Group X_5 as compared with the other conditions.

Evidence that performance will vary systematically with the number of repetitions of the training list, even if on only one of the originality tests, lends some further support to the interpretation of originality as a form of learned behavior obeying the principles of operant conditioning.

These results further highlight the failure in Exp. III to obtain any effect upon originality when unusual uses items were employed as the training materials. The fact that a single repetition of a 24-item training list produced a reliable increase in originality on both tests in the present experiment clearly indicates that the failure to obtain an effect in the earlier experiment was not due to an inadequate amount of training.

EXPERIMENT V

If, as we have assumed, originality can be learned according to the principles of instrumental conditioning, then it should show some degree of persistence following training. Since retention is one of the characteristics of learned behavior, the disposition to emit uncommon responses induced by the standard experimental training procedure should also show some degree of permanence. The purpose of the present study was to determine whether this in fact is the case.

Method

Subjects. The *S*s were 177 students drawn from introductory psychology classes.

Stimulus materials. The lists of stimulus words used in the initial free association training and subsequent tests are the same as those employed in the first three studies, and were administered in the same order as in Exp. II. As before, the Unusual Uses Test was administered after the free association test list.

Procedure. All *S*s were presented the initial free association training and final test lists of stimulus words in the same fashion. Two control groups received only a single presentation of the training list. Group C received the free association and Unusual Uses tests approximately one hour later. During the interval of time between the training and test sessions, *S*s were engaged in work on a series of 10 paper and pencil problems. Group C_2 was dismissed following the initial presentation of the free association training list with instructions to return in 48 hours. The free association and Unusual Uses tests were administered at that time. Two experimental conditions, X and X_2, experienced comparable delays between the initial and final test sessions. However, as in previous experiments, they received five additional presentations of the initial free association list with instructions to give a different response than the one used before to each stimulus word.

TABLE 17

MEAN PRETEST FREE ASSOCIATION ORIGINALITY SCORES IN EXP. V

Cond.	*N*[a]	Mean	*SD*
X	62	84.76	22.95
C	58	80.59	26.07
X_2	27	77.46	21.50
C_2	30	79.47	27.93

[a] Due to an unfortunate oversight, approximately half the *S*s in the two-hour delay groups were given a different form of the Unusual Uses Test which had the same frontispiece as the test previously used. We have therefore discarded the free association data from these *S*s as well as their unusual uses.

Results

Table 17 shows the mean originality scores obtained by the various groups on the initial presentation of the training list. A simple analysis of variance indicated that the groups did not differ significantly prior to the introduction of the experimental treatments ($F = .67$). Homogeneity of variance was indicated by the Hartley test.

Table 18 shows the mean originality scores obtained by the various groups on the

TABLE 18

UNADJUSTED AND ADJUSTED (MEAN′) FREE ASSOCIATION TEST ORIGINALITY SCORES IN EXP. V

Cond.	Mean	*SD*	Mean′
X	71.58	26.16	68.77
C	84.47	24.57	84.46
X_2	77.12	23.13	79.20
C_2	87.05	26.05	87.79

free association test of originality. Appropriate statistical tests indicated that the variances of the test scores were homogeneous as well as the regression of test scores on training scores.

Since the training and test scores were correlated significantly, a two-way analysis of covariance was conducted in which the originality test scores were adjusted for initial differences on the training list.

An *F* of 21.9 ($df = 1, 172$; $P < .001$) was obtained for the originality training effects. An *F* of 5.58 ($df = 1, 172$; $P < .05$) was obtained for the delay effect.

The frequency with which different uses were offered for each item in the Unusual Uses Test was determined on the basis of the norms from Exp. I, and originality and common response scores obtained as in the previous studies. Since the distributions of scores for unique responses were skewed, and the variances were heterogeneous, the transformation $\sqrt{X + .5}$ was applied to them.

Table 19 shows the mean number of transformed unique responses obtained by the four groups. Hartley's test indicated that the variances of the transformed scores were still somewhat heterogeneous, $P = .05$. As in previous studies, the measures of unique responses and common responses were not correlated significantly with originality on the initial free association list, thereby precluding the use of the analysis of covariance.

A two-way analysis of variance of the transformed unique responses yielded an *F* of 15.47 ($df = 1, 172$; $P < .001$) for the originality training effect. An insignificant *F* of 1.10 was obtained for the delay effect.

TABLE 19

MEAN NUMBER OF TRANSFORMED ($\sqrt{X + .5}$) UNIQUE USES IN EXP. V

Cond.	Mean	*SD*
X	1.42	.64
C	1.06	.45
X_2	1.49	.53
C_2	1.21	.57

TABLE 20

MEAN NUMBER OF COMMON USES IN EXP. V

Cond.	Mean	*SD*
X	18.06	4.84
C	18.15	5.24
X_2	17.74	5.37
C_2	15.67	3.97

Table 20 shows the mean number of common responses obtained by the four groups on the Unusual Uses Test. The variances were homogeneous. An *F* of .77 was obtained for the originality training effect, and 3.32 for the delay effect ($df = 1, 172$, in both cases). Neither effect was statistically significant. The results from the two different tests of originality indicate that highly reliable originality training effects do persist for some time. Thus, some additional support is gained for the interpretation of the training effects as a form of learning. Results from the free association originality test also show a significant delay effect or decrement in originality as the time between the initial and final tests is increased from one hour to two days. This, or course, is another common characteristic of learned behavior.

Results from the Unusual Uses Test showed a highly reliable training effect but not a delay effect. There was no significant decrease in the number of unique responses when the delay was increased from one hour to two days. Again, this discrepancy in the results obtained from the two originality tests may be due to a variety of reasons. It cannot be concluded, however, that a reliable decrement in unique uses did not occur within the time interval studied. Comparison of the delay conditions employed in this experiment with the standard experimental and control conditions of Exp. II, which received the free association lists in the same order, suggests that a decrement occurred within the first hour of delay.

The extent to which the effects of originality training persist, or are retained, is undoubtedly a function of many different variables of the sort that have been studied in classical experiments on verbal retention. Such variables as amount of practice or original learning, a wide range of delay intervals, activity during the delays, and distribution of practice trials, may all influence the extent to which the effects of originality training persists.

As data accumulate from parametric studies of this sort, and they show that performance on tests of originality vary in a fashion similar to performance in more traditional learning situations, to that extent it may be assumed that the effects of originality training do not differ fundamentally from other forms of learning. The study reported here is an initial attempt in that direction, and lends some support to the hypothesis that originality can be learned in the same fashion as other forms of operant behavior.

CONCLUSIONS

Each of the four experiments employing the standard experimental training procedure with free association materials produced a highly reliable increase in the uncommonness of responses on two different tests. The first two experiments suggest that the standard experimental training procedure of repeatedly evoking different responses to the same stimuli is the most successful of the three experimental procedures employed. It produced significantly greater facilitation of originality on one or both originality tests as compared with the method of evoking different responses by presenting different stimuli as in Exp. I, or evoking uncommon responses as textual responses as in Exp. II. Possible variables responsible for the efficacy of the standard training procedure have already been indicated earlier in this paper and elsewhere (Maltzman, 1960).

The results of Exp. IV and V lend some support to the hypothesis that originality is learned behavior and varies as a function of the same antecedent conditions as other forms of operant behavior. Effects of originality training may persist for as long as two days, under the training and test conditions employed in Exp. V. Originality, at least, on the free association test, also varies as a function of the number of repetitions of the training list.

Although the standard training procedure has proved to be effective, there are many possible variations of the procedure that remain to be investigated. For example, instead of completing the entire training list of stimulus words before re-presenting it, successive responses to each stimulus might be evoked. Various combinations of these procedures may also be studied. Further experimental studies are also needed employing different types of test criteria, including the more traditional kinds of problem situations. Studies of this kind are currently in progress in the UCLA laboratory. Work of this sort may indicate the extent to which the kind of training procedure studied here might have practical applicability. In this connection, it should be noted that the behavior measured on the originality tests employed in these experiments and similar tests correlates reliably with ratings of original behavior (Drevdahl, 1956).

In conclusion, it should be noted that the term "originality" as employed in this paper refers to a particular kind of behavior measured by specified operations under given conditions. If it is objected that this is not genuine originality, then it should be incumbent upon the critic to specify the latter behavior in equally operational terms so that it too may be subjected to experimental study.

REFERENCES

DREVDAHL, J. E. Factors of importance for creativity. *J. Clin. Psychol.*, 1956, **12**, 21–26.

GUILFORD, J. P. Creativity. *Amer. Psychologist*, 1950, **5**, 444–454.

JUDSON, A. J., & COFER, C. N. Reasoning as an associative process: I. "Direction" in a simple verbal problem. *Psychol. Rep.*, 1956, **2**, 469–476.

JUDSON, A. J., COFER, C. N., & GELFAND, S. Reasoning as an associative process: II. "Direction" in problem solving as a function of prior reinforcement of relevant responses. *Psychol. Rep.*, 1956, **2**, 501–507.

MALTZMAN, I. On the training of originality. *Psychol. Rev.*, 1960, **67**, in press.

MALTZMAN, I., BOGARTZ, W., & BREGER, L. A procedure for increasing word association originality and its transfer effects. *J. Exp. Psychol.*, 1958, **56**, 392–398.

MALTZMAN, I., BROOKS, L. O., BOGARTZ, W., & SUMMERS, S. S. The facilitation of problem solving by prior exposure to uncommon responses. *J. Exp. Psychol.*, 1958, **56**, 399–406.

SKINNER, B. F. *Verbal behavior.* New York: Appleton-Century-Crofts, 1957.

THORNDIKE, E. L., & LORGE, I. *The teacher's word book of 30,000 words.* New York: Bureau of Publications, Teachers College, 1944.

WILSON, D. P. An extension and evaluation of association word lists. Unpublished doctoral dissertation. Univer. of Southern California, 1942.

WOODWORTH, R. S., & WELLS, F. L. Association tests. *Psychol. Monogr.*, 1911, **13**, (5, Whole No. 57).

(Received September 8, 1959)

APPENDIX

THE CLASSIFICATION OF RESPONSES ON THE UNUSUAL USES TEST

The list of common and unique uses that follows was initially obtained from the results of Exp. I and was used with minor modifications in the scoring of the Unusual Uses Test from the later experiments. In the initial classification based on the results of Exp. I any use occurring only once in the distribution for all items was designated as unique. Uses occurring more than once were classified as common. Adjustments were made in the initial list by the addition of unique uses when they occurred in the test of a later experiment. On the other hand a unique use was reclassified as common if it occurred more than once on any given subsequent test. Few adjustments of this sort were necessary.

1. AN AUTOMOBILE TIRE (used on the wheel of an automobile)

Common Uses

float
roll
swing
bumper
toy
smoke
chair
obstacle course
jump on
scrap rubber
flower bed (planter)
weight
mending (patch) rubber
weapon (hit someone with, etc.)
decoration (garden, gas station, etc.)
shoe soles
target
advertisement
costume
ride on
inner tube (for swimming)
train dogs
melt and reshape
boat dock (bumper)
wheel on cart
slingshots
start fire (burn)
put in garage
blockade
football practices
insulation
rubber bands
throw things through (football, etc.)
sled
hoop
drum head
supporting another object
sell for living (retail product)
make smell
make sandals
make money
life saving
to swim with
binding to hold tree
horse equipment in horse show
play catch with
draw circles
buoy
tunnel
hold things in lining (hiding place, paper, water, etc.)
to throw
to cut up
make smog
sand box
to break things
hide narcotics
shock absorber

Unique Uses

prop up bigger object
holding hub cap
protect car from bumping paint
protect axle
lawn mower wheels
toss bean bags through
seat at ball game
have on fishing boat
cradle
base for lamp
cover
fort
door mats
anchor down poles
durability demonstration
cable holder
battering ram
padding on walls of amusement park
household rubber material
use in musical chairs
keep rug from slipping
hold air
test coordination
skin diver's net
hold letters or bundles
seat at beach
press papers
pillow
line doors or furniture
store refuse in
bumper cars at beach
bed for animal
hold door open
painter's still life
hold back loose dirt
keep tire salesmen in business
costume for strippers
picture frame
run in to build up legs
soap box racer brake
plug a hole
in a skit
make wall
base for bell-ringing pole at carnival
trade it
inner lining for rubber gun
hang things on
wedge
carve on
make rubber piece for trolling (fishing)
keep repair men busy
roof for stork nest
souvenir
ballast in a boat

- in relay games
- lift to build muscles
- trap
- rubber strings for tying wire
- hang on tree
- clutter up backyard
- for color scheme
- push things down driveway
- make tracks in sand
- use in circus act
- fill with cement for weight
- mold
- capture or hold someone
- build pyramid to jump off
- draw on with chalk
- build tree house
- sunbaths
- texture pieces in art projects
- loan to your neighbor
- put in a continental kit
- hold back flood waters
- build with
- throw knives
- makeshift table
- gift
- slice and make holders
- transfer air to bicycle
- sit on while being pulled by speedboat
- wagon
- to step on when there's mud
- to hide behind
- chewing gum (Spain)
- stick pins in
- steal
- wheel barrow
- in a clock
- adjust temperature
- on a safe
- rubber coverings
- support walnut trays above ground
- in zoo for monkey to hang on
- around a horse in a bull fight
- joke
- make mudpies in middle of
- play tag with
- tear up and use as brake on toys
- dogs to play with
- build fence
- end of hammer to pound out dents
- keep wood raft afloat
- tractors
- flatten something
- little boy to play army in
- burn to prevent frost damage
- store unused garden hose
- ash tray
- bird nest
- make an invention
- pile up outside of dealers
- cause accidents
- protective lining for mechanical clamp
- bottom of table legs to prevent harming floor

2. A KEY (used to open a lock)

Common Uses

- lever (pry things open)
- weight
- decoration (on chair, etc.)
- clean something (fingernail, crack, etc.)
- toy
- scratch (ear, etc.)
- screwdriver
- cutter
- noise
- charm
- melt for scrap metal
- start car
- pry open (jar, can, door, shoe polish, window, etc.)
- key to city
- scratch (message, desk, wood, paint, etc.)
- weapon
- symbol (religious, advertisement, etc.)
- necklace
- bracelet
- have (open) diary
- put on key ring or chain
- give to someone
- make new keys
- fish lure
- start mechanical toy
- bookmark
- hobby
- token
- wedge
- jingle (to sound like money)
- secret identification (password)
- rattle
- "index" for map, etc.
- dig (make) hole
- jewelry
- hypnotize
- tap code
- wind clock
- short circuit
- sorority badge (pin)
- tighten skate
- make blueprints
- use in hop-scotch
- look important
- to throw
- tool
- drawing outline
- carving
- hammer with
- pulley for window shade
- conductor
- treasure in treasure hunt
- make designs in sand
- magnet
- gift
- poke hole (in cloth, wood, etc.)
- in game of hide and seek
- paperweight
- clue
- wear
- souvenir (to hang on wall)
- diving and retrieving

Unique Uses

- add to junk box
- plier
- chrome plating
- electric shock
- having secret
- point out something
- scrape rust on car chrome
- ornament on Christmas tree
- symbolize gift of car
- signify two people going together
- button
- hoop pick (for horses)
- shoot in slingshot
- tap rhythmically
- fuse
- impress people
- studying designs

straighten hairpins
lift small objects
give to baby to chew on
chew on if anxious
to write with
door knocker
award
unfasten screen
school remembrance
scrape
to press
paint it gold
turn on furnace
thermometer on wall
engrave in tree
make print
pad a lock
wrap string around
put on ribbon and tie around neck
to photograph
cut teeth with
press down bump
design on pictures
weight for small animals
impression in clay
evidence in trials
training dogs
toothpick
brace
break something
tap on windows or doors
place marker
play catch with
fiddle with
honorary club
antique
jangle to attract attention
reflect sunlight
suck on
put in valentine
put in scrap book for memory
object for conversation
put safety pin in hole and save
decorate plastic panels
hat
melt into buttons
handle on box
melt and make ring
loosen something
twirl
pick ear with
arouse curiosity

3. A SAFETY PIN (used for fastening)

Common Uses

prick skin
punch hole
open lock
decoration
scratch
weapon
fish lure
remove splinter
pin
toy
toothpick
sticking to bulletin board
put things on map
stick someone
make chains
untie knots
puncture blister
break balloons
remove sliver from skin
decorate sweaters
hold diaper
weights
clean fingernails
open glue tube
punch holes in (cut) paper
carve words
open objects
clean small objects
insert elastic (in sleeve, belt, object)
necklace
pry things open
hold other pins
sewing
tack (something to wall)
picture hanger
marker for hopscotch
jewelry
puncture tire
"pick" (poke) at something
hold dress (skirt) together
hold papers together
pull cord through opening
use as thumb tack
clean small ledges (corners, etc.)
repair zipper
key ring
stick something
retail product
put on button with
pin paper to wall
decoration for hats or dresses (clothes)
kill insect (bug)
get dirt out of ring
engraving (carving)
make wound (sore)
mark stitches in knitting
make wire (scrap)
hang up skirt
hang things on wall
patterns or designs
clothes pin
screwdriver
repair something
scratch name in metal
indicate markers
pick things apart
scratch furniture (or other objects)
stick pin in cushion
hold buttons
letter opener

Unique Uses

connect as belt
base for decorative pin
measure acoustics
make doorbells stick
to keep in pocket
repair hems
scrap metal
cork opener
make link in S. P. bracelet
demonstrate magnate
make fine lines in drawing
hook for lanyard
squeezer
decoration for food
string tokens on
hold glasses together
prop
poke things that are wedged in places
string pearls
pull string through narrow opening
hold message on door
keep window shade up

mark spot on treasure hunt
wire joiner
pierced earrings
prick babies
to impair clothes washer mechanism
hold up falling objects
determine who has next baby at shower
pinpoint object
cause pants to fall down
poison dart
carry in purse
get money out of piggy bank
clean top of salt shaker
bend into party favor
widen small hole
play "pin the tail on the donkey"
phonograph needle
puncture leather
collar clip
pick off fingernail polish
bracelet
Christmas tree decoration
belt
pinch somebody
bend around something
use as nail
torture flies
stick in a flower
make small noise
running elastic
clean pipe
on back of sign
clean comb
make scatter pin
make up weight
terms of barter
make cardboard filter
reshape
tear material
remove thread
stopper holes
mechanical purposes
separate stitches in knitting
dig hole
stick butterfly on it
play poker with

4. A WATCH (used for telling time)

Common Uses

decoration
toy
demonstrate mechanics
(paper) weight
tell direction
hear (listen to tick)
pawn (sell)
prize
hypnotize
rhythm
bracelet
for experiments
impress people
draw circles
timing events
take apart and fix
jewelry
gold content
picture holder
used by watchmakers (to make living from)
cover scars on wrist
suspend fobs on
gift
art pieces
tell date
time pulse
to wear
wealth object
showmanship (show off)
learn to be jeweler
take apart and study parts
test hearing
time tests
teach how to tell time
time bomb (prank)
bribe object
to start conversation with
to keep a date (appointment)
timing device for bombs
use as pin
set off alarm
put cheat notes in
teach to count (numbers)
check resistance to water
noise to keep someone awake
advertising
throw at someone
metronome
to sell
tell weather
(hang from) chain
ticking sound to put to sleep
smuggle (conceal) (jewels)
collect as hobby
family heirloom
souvenir
to listen to
to annoy somebody
strap things together
ticking to soothe puppy
noise from ticking
comfort a (motherless) cat
looks
loan to magician who pretends (tricks) to break it.

Unique Uses

melt parts for metal
money
learn clockwise vs. counterclockwise
make dent in arm
reading meters
glass for cutting
irritate person
silver content
reassemble gears for other purpose
provide employment
security for loan
fiddle with when nervous
hang on string and find vertical
use chain to latch door
use parts—paraffin for wrapping paper
speedometers
to steal
glass for magnifying
check resistance to shock
exercise fingers
test sight
home appliance
guessing lengths of time
mechanics to fool with
lure to catch thief
swing in self-defense
knuckle guards in fight
to expand industry
forecast future
symbol
clean fingernail with stem
training item in school

projectile for throwing
investment
window displays
reward
split apart for modern abstraction
mantelpiece
tinkering
wrist warmer
source of frustration when not running
starting signal
pounding something
vent destructive instincts on it
back as mirror
hold something on wrist
get spare parts
open something
teach someone how to make one
weight measure
parts to decorate fish bowl
squash ants
sound effect in movie
keeps jewelers happy
hide birthmark
threatening someone
practice in amplifying sound
turn on lights
turn on pump
hide stolen article
puzzle
remembrance
give seasons
symmetrically suspended center piece
chocolate candy formed as watch
sound vibration experiment
used to tell when you are near magnet
break and use parts for confetti
practice putting things together
reassure baby chicks
weight on tie to keep it in place
guard a vault
hang medal from
wear around neck to hold collar in place
play passing game
decorate with parts in plastic panels
take out springs and bounce them around
buy bands for
spend money fixing
in place of grandfather clock
take hands off and use in another watch
drop from high distance to test breakage

5. A BUTTON (used to fasten things)

Common Uses

decoration (shoes, handbags, etc.)
toy (play game)
play money
draw circle
necklace
earrings
poker chips (gambling)
means of exchange
make objects (picture, etc.)
stop up sink
hobby
teach how to count
practice sewing
to throw (at people)
lucky pieces
badge (emblem)
checkers
bingo
make rattles
noses for dolls, etc.
make designs (patterns, etc.)
sell for money
make eyes (on dolls, on animals, snowmen, etc.)
learn colors from
slug
small target
to suck (chew) on
monocle
trade object
bracelet
teach children to fasten things (dress themselves)
make noise (with many)
play with when nervous
make jewelry
fill button (sewing) box
"ivy league look"
hide tear or spot (on dress)
hide a hole (in wall, etc.)
in counters
put string through it (bends, etc.)
hold something together
advertise for office in election

Unique Uses

pry open cans, jars, etc.
to roll
souvenirs
paperweight
color board
for beauty or looks
swing from swing to scare flies
pressure object to stop bleeding
golf marker
article in visual demonstration
map marker
to paint
stuff chair
pin
construct pendulum
to flip as coin
put texture on paintings
use up thread
demonstrate durability
signal
teething for babies
weight
hold name tag
stop swelling of bump
Christmas tree decoration
separate papers (wedge, etc.)
prize
historical value
teach handicapped children self-help
sift bread crumbs through holes
stuff dolls
make faces with
shimmy
stuff bean bag
mark of crosswalk
word in word assimilation tests
memories of clothes
for art projects
hang from string to see where wind blows
outline pictures
clue
swallow to get sick

- put in fish pond
- evidence in a trial
- in educational books for children
- trip someone
- build houses
- end of string attached to light
- play catch with
- knob on radio
- token of appreciation
- put in muffins for raisins
- fingernails on snowman
- remembrance
- put in goulash for joke
- throw in shallow pond for color
- outlet for creative design
- nervous fingers to fidget with
- fishing plug
- wheels of a car
- show authority
- to trip on
- designate political party
- button box for invalids
- make clip for hair
- make knob on lid
- eyes in gingerbread cookies
- pledge girlfriend with
- hold the car door in place
- causes embarrassment when they fall off
- toy soldiers
- put on string for cat to play with
- make pools in a lake

6. EYEGLASSES (used to improve vision)

Common Uses

- change appearance (looks)
- protect eyes (from glares, sparks, etc.)
- magnify
- disguise (hide behind, etc.)
- acting (plays)
- cutting
- keep out of fights
- toy
- decoration
- hearing aids
- keep out sun
- look intelligent (professional, etc.)
- for reading (studying, etc.)
- start fires
- prop
- masquerade costume
- comedy
- when you want to break something (release emotions)
- make eyes look larger (or smaller)
- put into a case
- sophistication (look distinguished)
- look distinctive or old
- telescope (binoculars, etc.)
- microscope
- mirror
- gesture object
- worsen vision
- give false impression (impersonations)
- hurt looks
- place on head to hold hair down
- protect eyes from smog (dust, etc.)
- relieve headaches
- fashion mode
- art object
- "gag gift"
- hide black eye
- rest eyes
- lenses for experiment
- paperweight
- weapon
- jewelry
- chew on ends (during exams, etc.)
- impress people
- for amusement
- carry in purse

Unique Uses

- shield metal particles
- pane for peephole
- doll house window
- bottle caps
- put pencil in
- tap to be annoying
- give self-confidence
- as fad
- test vision
- look at atom test
- crush glass and use as dog poison
- spend parents' money
- decorate with shells
- attract attention
- learn to use radius of curvature
- give people complex
- scare away men
- give to charity
- sun reflector to send signals
- hold false nose on
- frame for spare wire
- add color
- spend money foolishly
- hang on statues
- to finger when nervous
- excuse for not obeying signs
- design frame for
- construct them
- melt down glass for better use
- scratch head with rims
- draw circle
- melt and analyze composition
- rinse the eyes
- replace human eye
- book mark
- conversation piece
- reach object that has fallen in narrow hole
- advertising
- cut oneself
- make living from
- depict certain personality
- to break and cost money
- hide hangover

Vol. 74, No. 7 Whole No. 494, 1960

Psychological Monographs: General and Applied

PERSONALITY FUNCTIONS OF SYMBOLIC SEXUAL AROUSAL TO MUSIC[1]

MICHAEL A. WALLACH[2] AND CAROL GREENBERG

Harvard University and the Age Center of New England, Inc.

How does music psychologically affect the listener? In most of the research to date done with the aim of answering this question, evaluation of psychological effects has depended on overt judgments by the listener concerning the music's emotional meaning for him. This tradition of studies has been recently summarized by Farnsworth (1958), and work by Sopchak (1955, 1957) provides a current example of the technique. Its use seems based on the historical linkage of experimental aesthetics with the introspective method, and would appear to neglect our knowledge of the inaccuracy often involved in conscious reports of motivational states.

Freud (1938a) has suggested that art in general can serve as a way of sublimating sexual energy, thereby channeling it away from its more customary social expression. Music, in these terms, could function as an indirect, compensatory means of expressing sexual arousal. Sublimation's relevance to music is well expressed in Erikson's (1958) description of Martin Luther:

> Most of all, this kind of person must shy away from intimacy. Any physical closeness, with either sex, arouses at the same time both an impulse to merge with the other person and a fear of losing autonomy and individuation. . . . The contrast between the exalted sexual fusion of his autoerotic dreams and the complete sense of isolation in the presence of the other sex is catastrophic. . . . Music, at such times, can be a very important means of socializing and yet of communing with one's emotions: and Martin was a master in using music as a bridge to others, and also as a means of creating distance (pp. 101–102).

So also, Sullivan (1949) repeatedly illustrates how Beethoven's emotional involvement with music performed a displacement function to compensate for his sexual withdrawal from other people.

Insofar as the motivational functions of music may carry sexual meanings, we can hope to understand them only through exploration at a more unconscious level, for two basic reasons. Even with the fullest intention of being candid, a person may be unable to report veridically on his sexual feelings because of their deep-lying connections with early fears. Secondly, societal conventions are such as strongly to inhibit such reporting even if a person actually is aware of these feelings.

Erikson's implied analogy between the psychological functions of dreams and those of music suggests, further, the following way of assessing music's unconscious motivational effects. The kinds of symbolism studied by Freud (1938b), in order to comprehend the latent sexual significance of dreams, may also help us to evaluate the possible sexual import of music. Toward this end, the study to be reported used variations of techniques developed by Clark (1955) and Beardslee and Fogelson (1958) as its point of departure. In brief, a listener was asked to let himself get into the mood of the music, and was told that after hearing the selection he would be requested to write a four-minute story which the music had suggested to him. Content analysis of the story for activities and things presumably symbolic of sexual arousal (in terms of Freudian assumptions) then permitted an evaluation of that selection's implicit sexual motivation effect for the listener in question.

---

[1] This study was supported in part by a research grant (M-2269) to the senior author from the National Institute of Mental Health, United States Public Health Service, conducted under the auspices of the Age Center of New England, Inc.; and in part by a grant from the Laboratory of Social Relations, Harvard University. Grateful thanks are due the members of the senior author's Harvard University course and seminar on the psychological foundations of aesthetics, for aid with various aspects of the study; and Donald S. Dunbar of Simmons College, for making classes available. Lise Wallach made invaluable critical comments on the manuscript, and drafts were also read by Albert J. Caron, Ellen Greenberger, and Nathan Kogan.

[2] Now at Massachusetts Institute of Technology.

The aim of the present research was to determine personality dimensions responsible for the effects music might exercise on symbolically expressed sexual motivation. The experiment's starting point was a psychological theory of aesthetics presented elsewhere (Wallach, 1959), which calls for a functional analysis of music—a study of what music accomplishes for the organism in terms of its needs. Our earlier considerations have suggested that music can function as a channel for the displacement of sexual motivation inhibited from more direct expression. What are the conditions of personality under which music will serve this function for the listener?

Clinical discussions of the reasons for displacement of sexual drive to fantasy levels (for example, Fenichel, 1945) suggest that such indirect fulfillment becomes necessary when, as in the case of Erikson's description of Luther, the individual feels a need for other people, but is afraid of them—by virtue of previous punishments or their threat. He therefore avoids closeness with others, but the need is sufficiently strong that his defense of social isolation cannot really work. The sign that such a person's defense of avowed introversion doesn't function successfully is the presence of a high level of anxiety.

Consider, on the other hand, someone who likewise professes a desire to be alone, but does not in addition manifest a high degree of anxiety. Such a lack of anxiety would seem to imply that this person's attitude of avoiding others is a much more stable adaptation. This individual, we would expect, hence would not need to use music as a source of substitute satisfaction in the manner just described for the introvert who is high in anxiety level.

Finally, we might well expect a person who avows a high degree of concern with other people—a social extravert—also to reflect this involvement to some extent on the fantasy level because of that involvement's very intensity. Such fantasy expression should be weaker than that energized by a displacement function, however, since in the latter case fantasy expression is the only channel available to the individual, while a social extravert is also able to express himself more directly. Since introversion seems to be a more "defensive" and extraversion a more "normal" orientation, anxiety more clearly seems to be an index of the success or failure of a defense in the case of introversion than in the case of extraversion.

More fantasy expression of sexual needs should occur for a social introvert, then, if the person has a high anxiety level, thus betraying a lack of success for his defense of social isolation. The result should be fantasy compensation. For introverts whose anxiety level is low, on the other hand, the apparent success of the introversion defense precludes the need for their engaging in fantasy displacement. Since social extraverts, finally, should show some symbolic sexual arousal simply as a reflection of their involvement with others, we would expect them to fall between the two introvert groups in level of symbolic sexual arousal.

In sum, the theoretical proposals we have been advancing are these:

1. One of the functions of listening to music is to permit the symbolic expression of sexual impulses.

2. The strongest impetus toward such expression arises for individuals who claim that they want nothing to do with others, but whose attitude of social introversion is unsuccessful as a mode of defense, as indicated by the presence of a high level of anxiety. Since direct expression of sexual impulses is thereby prevented but the need for others still remains, the only available channel is one of displacement.

3. A weaker impetus toward such expression arises for individuals who describe themselves as socially extraverted. Since direct expression of sexual impulses is not prevented by the person's attitude, the function of symbolic expression is one of reflecting their involvement with others,

rather than of displacement. Not being the only channel of sexual expression, the intensity of symbolic arousal should be less than in the first case.

4. The weakest impetus toward such expression is present for individuals whose expressed defense of social introversion is successful, as indicated by the absence of anxiety, since its success implies that the need for other people has been conquered. Hence, a minimum of symbolic expression is required.

The following hypotheses may be derived from the above propositions, and constitute the experimental questions of the present study:

1. Among those high in admitted anxiety level, subjects calling themselves socially introverted will show a *greater* degree of symbolic sexual arousal to music than subjects calling themselves socially extraverted.

2. Among those low in admitted anxiety level, subjects calling themselves socially introverted will show a *lesser* degree of symbolic sexual arousal to music than subjects calling themselves socially extraverted.

METHOD

Selection of Music and Independent Judgments of Arousal Value

Three jazz selections were played to a Harvard-Radcliffe class of 10 male and 13 female undergraduates who were asked to evaluate them in terms of sexual arousal. After hearing each piece, they were to rate it on a seven-point absolute scale of arousal value ranging from "most arousing" at one end, to "least arousing" at the other. A derivation of judges' rank orderings was made from the position of each of the three pieces on this scale. Rank order rather than absolute scores were used in order to control for differences among judges in the position of all three pieces in absolute arousal level.

The significance of the mean difference between ranks given each pair of pieces was ascertained for each sex separately and both combined, the measure in all cases being an individual's rank difference for two selections. On the basis of this analysis, reported in detail below, it was found that one of the three pieces was judged lower in sexual arousal value than the others, while the latter two were not distinguishable in that regard. The order of presentation of these pieces was decided accordingly, the piece that was lower in judged arousal value being played first. The aim of this order was to preclude any possible halo effect arising from the transfer of a high arousal piece's imagery to a subsequently heard low arousal selection.[3]

The selections heard, in their order of presentation, were Eddie Burt's "Falling in Love," Dizzy Gillespie's "Doodlin'" (these two pieces each fading out right before the restatement of the theme), and Sam Taylor's "Sam's Blues." The lengths of these selections were 185, 147, and 146 seconds, respectively, and they were played from a high fidelity tape.[4]

Subjects

A total of 98 undergraduate women participated in the story session part of the study—students in introductory psychology classes at Simmons College. The sample was restricted to women because Beardslee and Fogelson's work indicated women to be more open to symbolic sexual arousal by music than men, a result whose meaning can be better understood in the light of our own findings below. For two reasons, however, the actual sample considered in data analyses was always smaller than 98. First of all, three class sessions were required to gather the data. Since these sessions were separated by intervals of about a month, different absences occurred in each, thus reducing the effective sample size to those who were present during all three hours for purposes of certain analyses, or during two of the three hours for purposes of others. A second cause of reduced sample size was the need to equalize cell membership in the two-way analyses of variance reported later. This required random elimination of cases from the oversize cells.

Story Session

During the first class session, the subjects (*S*s) wrote stories in response to the musical selections. At the outset they heard the following instructions:

> This is a test of creative imagination. It gives you an opportunity to use your imagination, to show how you can create things as suggested to you by music that you hear. You are going to hear several musical selections. Try to let the mood of each one come over you completely—just feel whatever the music makes you feel; get completely involved in it—lose yourself completely in the music.
>
> In a minute I will play the first selection. Remember, just lose yourself completely in the music—let its mood really come over you. After you hear the music you will be asked to write a story suggested to you by this music. The story

---

[3] The pieces were also in this order when heard by the independent judges.

[4] Their respective record numbers are Jazztone 1223, Verve MG-V-8242, and MGM E-293.

> can be of any kind, but it should be suggested to you by the music you have just heard. You may write notes during the music if you want to, but don't begin to write the story itself until the music is over. Remember, this is a test of creative imagination. We will play the first selection now; try to lose yourself completely in it. When it is over you will start to write the first story.

After hearing the piece, the following was read:

> Now you have four minutes in which to write a story suggested to you by the music that you have just heard. Please write it on the first page of your booklet.

Instructions concerning the second and third selections were abridged but otherwise similar forms of the passages just presented. Each story was written on a separate page of the test booklet, *S* having a blank page in front of her while listening to a piece. The stories obtained were scored for symbolic sexual arousal in a manner described below.

Preference Session

A second class administration occurred about a month later. Its purpose was to obtain a judgment of absolute level of preference for each of the pieces which had been heard before. Absolute preference levels were needed in order to control for the possibility that relations of personality to arousal would occur only for people who generally liked music of this kind.

The *S*s received a sheet with three seven-point scales labeled "A," "B," and "C," respectively; the left end of each scale having "Strongly Dislike" printed above it, and the right end, "Strongly Like." This preference session was introduced as follows:

> I am going to play several musical selections which you have already heard. After you listen to each one, I want you to tell me how much you like or dislike it, by circling one of the numbers along the row. If you dislike the piece very much, you would circle 1; if you like it very much, you would circle 7; if you were indifferent to it, you would circle 4. And so on for the other numbers: if you liked it a little, you would circle 5; if you liked it quite a bit, circle 6; and if you liked it very much, circle 7. On the other hand, if you disliked it a little, circle 3; if you disliked it quite a bit, circle 2; and if you disliked it very much, circle 1.
>
> The first musical selection will be called "A," so you circle a number on the first line. The second piece is "B," so you mark the second line for that piece; and, likewise, the third piece is "C."
>
> Don't feel that you should express any particular kinds of likes or dislikes, since there obviously are no right or wrong answers. Some of you may tend to like all the pieces; some of you may tend to dislike all of them; still others of you may like one piece more, dislike another one, and so on. Just be sure that each of your markings expresses your true feeling about the piece you just heard.
>
> Now you will hear Selection A.

After listening to this selection, *S* was asked to mark how she felt about Piece A. This process was then repeated for the second and third pieces. The same order of presentation for the pieces was used here as in the story session.

Personality Assessment Session

The hypotheses stated earlier required us to assess two dimensions of personality: the individual's level of admitted or manifest anxiety and her professed attitude of social extraversion or introversion. Examination of Eysenck's Maudsley Personality Inventory (Eysenck, 1956; Jensen, 1958) indicated that it provided highly suitable ways of measuring each. The test contains two scales used in the present study. The Neuroticism scale, containing such items as "Are you inclined to be moody?" was found to correlate highly with the Taylor Manifest Anxiety Scale (1953), and hence provides an index of the extent to which an individual believes himself to be anxious. The Extraversion scale, containing such items as "Would you be very unhappy if you were prevented from making numerous social contacts?" correlated highly with the Social Extraversion scale of the Minnesota T-S-E Inventory (Evans & McConnell, 1957), the correlation being much smaller with that test's Emotional Extraversion scale. Inspection of the Maudsley Inventory's extraversion items showed them clearly to concern the social rather than the emotional forms of extraversion-introversion—i.e., to concern desire for the company of others rather than avowed openness of emotional expression.

While our hypotheses specifically concerned social extraversion, it was of interest to determine the extent to which emotional and social extraversion in fact differed. An index of emotional extraversion, therefore, was also added to our test battery, in order that both kinds of extraversion would be measured. The Emotional Extraversion scale from the Minnesota T-S-E was used for this purpose, its items being randomly interspersed among the manifest anxiety and social extraversion items of the Maudsley test. The Eysenck Extraversion scale was preferred to the Social Extraversion scale from the T-S-E, because the former was developed more recently and its relations with other instruments more thoroughly investigated. The whole untitled questionnaire was administered during a third class hour about a month after the preference session. The Neuroticism and Extraversion scales of the Maudsley test and the Emotional Extraversion scale of the Minnesota T-S-E were scored in the ways described by these tests' respective authors.

*S*s identified themselves by name on the procedures used in each of the three class periods.

Scoring Categories for the Stories

Each story was scored for presence or absence of five variables. These variables may be briefly described as follows:[5]

1. *Motion*. This variable was scored if the story described movement by the body as a whole or by any part of the body from the shoulders down, hence excluding head, arms, and hands. The activity of dancing, however, was not scored.

2. *Rhythm*. Rhythm was scored for any repetitive action, sound, or adverb phrase such as "back and forth." Verbs concerning progressive movement over space, such as "dancing" or "walking," were scored for rhythm only if a modifier was added indicating that the movement was done with rhythmic repetition, such as "he walked with even paces."

3. *Peak*. A story was scored for peak if there was a rise followed by a decline in the level of tension implied by the action.

4. *Penetration*. This category was scored for imagery concerning forceful pushing through a resistant medium or one which was treated as resistant in the story.

5. *Symbolic content*. The definition of sexually symbolic content was based on Gutheil's formalization (1951, Ch. 3) of Freudian dream theory (Freud, 1938b). With the exception of dancing, an item of story content was considered sexually symbolic if it appeared on Gutheil's list, and the scorer was asked to decide on presence or absence of symbolic content in the story.

The view that these variables concern symbolic sexual arousal depends on symbol interpretation assumptions made initially by Freud. Strictly speaking, the link from our analysis of the stories to symbolic sexual arousal is assumed, not tested, by the present study. It has been experimentally confirmed, however, in other studies, of which we may cite, for instance, the work of Clark (1955). One could, on the other hand, avoid this assumption, simply consider our scoring categories as indicative of some sort of motivational arousal, and interpret the relationships obtained between personality and arousal from this starting point. In particular, one might provide an interpretation in terms of such a motivational dimension as emotionality–rationality, or subjectiveness–objectiveness. No evidence within the present research overrules this possibility. But since such an approach would render the obtained relationships hard to understand, and indeed would not have generated the theoretical propositions which formed the impetus for our hypotheses and experiment, we wonder at its advisability here; especially since the assumption of symbolic sexual arousal is rendered tenable by such independent experimental evidence as the study just mentioned.

Two points internal to the present experiment provide further support for the assumption. The symbolic content category is very hard to explain in terms other than sexual symbolism, and its trends across musical selections are consistent with those of all four remaining scoring categories (as presented in detail in the following section). The relevance of these five variables to sexual arousal is also indicated by the general correspondence between the scores they yield for the sexual arousal value of each piece and independent explicit judgments of these pieces' relative value for such arousal (also presented in detail below).

Since there were three stories per person, each person's total score could vary from zero (absence of all five variables in all three stories) to 15 (presence of all five variables in all three stories), the possible score for a single story ranging from zero to five. All the stories were scored "blindly" by two independent scorers. As the total number of *S*s writing stories was 98, there were 294 stories in all. The stories were assigned the numbers one to 294 in a random order, each story appearing on a separate piece of paper and being identified to the scorer only by this number. A scorer thus had no idea which stories were written by the same individual, nor which stories were written to each of the three musical selections. A number code indicating which three randomly assigned numbers corresponded to the three successive stories for each individual permitted the eventual ordering of story scores in terms of individuals and musical pieces.

RESULTS AND DISCUSSION

Independent Judgments of Arousal Value

Table 1 presents the rank ordering of our three musical selections for sexual arousal value, using a sample of Harvard-Radcliffe undergraduates as judges. The analysis in columns four to six is based on the mean difference between the ranks given each pair of pieces, the two-tailed probability level in each case indicating the likelihood of obtaining this great a deviation

---

[5] Scoring categories for the first four variables were taken from Beardslee and Fogelson (1958), except for the decision not to score dancing for the motion variable. This decision was based on the generally widespread appearance of dancing imagery in all stories, due, apparently, to direct associations to the nightclub or social dance milieu in which jazz typically is heard. Scoring for the fifth variable was derived from Clark (1955), with two changes: instead of scoring for strong presence, some presence, or absence of symbolic sexual content, a two-category judgment of presence or absence was made. Secondly, content concerning dancing was not scored, for the reason just described. More detailed descriptions of the scoring categories than those presented are available upon request from the senior author.

TABLE 1

INDEPENDENT JUDGES' MEAN RANKINGS OF THE MUSICAL SELECTIONS FOR SEXUAL AROUSAL VALUE

Selection	Mean Rank			Test of Mean Difference between Ranks					
				B-A		C-A		C-B	
	A	B	C	t	$p <$	t	$p <$	t	$p \vee$
Men ($N = 10$)	2.8	1.9	1.3	2.62	.05	9.06	.001	1.60	*ns*
Women ($N = 13$)	2.6	1.7	1.7	2.76	.02	2.37	.05	0.00	*ns*
Combined ($N = 23$)	2.7	1.8	1.5	3.64	.01	5.11	.001	1.04	*ns*

Note.—A rank of 1 indicates highest arousal value; a rank of 3 indicates lowest arousal value.

from zero by chance alone. The mean ranks in columns one to three are presented to show the magnitude of the differences, but are not used in further calculations since the appropriate statistical test requires the use of rank differences for individuals. It is apparent that both men and women make similar arousal value judgments: for each sex, Selection A being considered least arousing, with B and C higher and not differentiated in arousal value.

Reliability of Story Scoring

Scoring reliability was determined on the basis of the largest possible sample—namely, all 98 *S*s who had the story session. Two scorers independently rated the materials on each of the five scoring variables discussed above.[6] As each *S* wrote three stories, and as the score for each variable consisted in a judgment of presence or absence, a scorer made a presence or absence rating for 294 stories in the case of each variable. The measure of reliability in scoring any one variable was the number of stories for which both scorers agreed in making a presence or absence decision divided by the total number of stories, or 294. This quotient, expressed as a percentage of the total, was 84% for motion, 88% for rhythm, 88% for peak, 98% for penetration, and 76% for symbolic content. The average reliability of scoring for all five variables combined was 86.8%. One of the scorers resolved those cases where disagreement had occurred.

Stimulus Value of the Music and Arousal

To what extent did the implicit measure of arousal obtained from the story content of our sample of 98 women reflect the independent judges' explicit decisions as to the selections' comparative arousal values? Table 2 presents the findings separately for each of the five scoring variables as well as for all combined. The means of the first three columns indicate that for each of the variables, and especially for all of them taken together, symbolic sexual arousal was lowest for Selection A, highest for Piece B, and in between for Piece C.

Are these trends reliable? The answer was provided in a manner analogous to that used for the independent judges described earlier. For each variable alone, and for all combined, the difference was determined between an individual's scores for each two selections. A *t* test then indicated whether this difference was significantly deviant from zero. From the obtained *t* values, presented in columns four to six of Table 2, we find the arousal difference between

---

[6] It might be noted that one scorer was male and the other female.

Pieces B and A significant beyond the .01 or .001 levels for four of the five scoring variables and for all combined. The difference in arousal between Pieces C and A passes the .02 level for two variables and the .001 level for a third, also being beyond the .001 level for the combination of all five. Finally, while the difference between Selections C and B is significant for only one of the five variables taken separately, its significance level for that variable is well beyond .01, and the difference for all variables combined passes the .05 level.

The statistical evidence hence suggests that the order of increasing arousal value for these selections is A, C, B. While Piece A was also placed at the low extreme of arousal value by our independent judges, they did not make a discrimination between the two remaining selections but rather considered them about equally high. On the whole, however, the match was quite good between these explicit and implicit measures of the sexual arousal of musical selections obtained from independent samples of *S*s.

In sum, the results just presented indicate that the music is not incidental to the arousal content found in our story productions. Rather, such arousal reflects independently made explicit judgments as to the differing sexual arousal value of the selections. Finally, no cumulative series effect seems able to account for the differing degrees of arousal content for stories written to the three pieces, since the arousal level for the selection heard last was lower than that for the preceding selection.

Techniques of Analysis for Personality Effects

Before turning to the findings on personality, some general considerations are necessary concerning the analysis techniques used. All the two-way analyses of variance reported, whatever the dependent variable in question, involved the use of some two of the following four dimensions as the independent variables whose effects were to be tested: manifest anxiety level, social extraversion–introversion, emotional extraversion–introversion, and absolute preference level. Each analysis required the construction of four groups representing the following combinations of any two of these independent variables: groups that were high–high, high–low, low–high, and low–low. Accordingly, each variable was dichotomized by splitting the sample into highs and lows, the split being as close to the median as possible. The split could never be exactly at the median because, in the case of each variable, there were a number of *S*s tied for a score that overlapped the median. In all four cases the overlap around the median was asymmetrical: that is, most individuals having the score value

TABLE 2

Experimental Sample's Mean Arousal Scores for the Musical Selections
($N = 98$)

Selection	Mean Arousal Score			Test of Mean Difference between Arousal Scores					
				B-A		C-A		C-B	
	A	B	C	*t*	*p*<	*t*	*p*<	*t*	*p*<
Motion	0.26	0.49	0.45	3.45	.001	2.50	.02	0.86	*ns*
Rhythm	0.13	0.33	0.15	3.36	.01	0.43	*ns*	3.07	.01
Peak	0.08	0.23	0.21	3.37	.01	2.43	.02	0.35	*ns*
Penetration	0.01	0.06	0.05	1.52	*ns*	1.65	*ns*	0.38	*ns*
Symbolic Content	0.24	0.58	0.48	4.71	.001	4.06	.001	1.29	*ns*
Total	0.72	1.69	1.34	6.22	.001	4.09	.001	2.49	.05

in question fell on one side of the median and only a few on the other.

Because of this asymmetrical nature of the overlap, and because of the other factors already mentioned which necessarily resulted in reduction of sample size, it was inadvisable to drop a whole group of cases just because they all had a score value which fell at the median of a distribution. The groups were dichotomized so as to yield as nearly equal halves as possible, the actual resultant cutting point for highs and lows deviating no more than 5% from the ideal of a 50%-50% split. That is, considering the four variables that were dichotomized, the extent of the cutting point's variation from half-and-half never exceeded 45%-55% break, and indeed was closer to half-and-half than that in all but one of the four splits made.[7] Thus, the four groups examined in any analysis of variance represented, for the two independent variables of the analysis, those individuals above the cutting point for both, below it for both, and above it for one and below it for the other.

Since two-way analysis of variance with more than one S per cell is very difficult unless the numbers of Ss in the various cells are equal, the cell size for an analysis was taken as the smallest number of Ss falling in any one of the four cells. Any Ss in excess of that number were removed from the remaining three cells by a process of random elimination. The variances of the four groups constituting each two-by-two table described were found to be homogeneous by Bartlett's test (Snedecor, 1946, pp. 249–250), thus meeting the assumptions of analysis of variance technique in all cases. Finally, the fixed constants model (McNemar, 1955, pp. 304–306) was used in the analyses, since none of the bases of classification involved any random sampling. According to this model, the within-cells mean square is used as the error term in calculating F ratios for rows, columns, and interaction effects.

TABLE 3

MEAN LEVELS OF AROUSAL TO MUSIC FOR ANXIETY AND PREFERENCE GROUPS

($N = 13$ per cell)

	High Anxiety	Low Anxiety
High Preference	5.08	3.62
Low Preference	5.46	3.38

Manifest Anxiety Level and Social Extraversion–Introversion

We turn now to the experiment's major concern. Our hypotheses, it will be recalled, postulated an interactive effect of manifest anxiety and social extraversion on symbolic sexual arousability to music. Before determining whether this interaction occurred, however, it was necessary to investigate whether differences in absolute levels of preference for the jazz music as a whole might act to influence the relationship between these personality dimensions and arousal to such music. Accordingly, differential arousability first had to be assessed separately in terms not only of an individual's position on each personality dimension, but also in terms of whether she tended, on the whole, to be high or low in her liking for the jazz music heard. A personality effect might conceivably be demonstrable for individuals liking the music but not for those disliking it.

Tables 3 and 4 present the means and analysis of variance for those high and low in manifest anxiety level, the Ss also being

TABLE 4

ANALYSIS OF VARIANCE, AROUSAL TO MUSIC FOR ANXIETY AND PREFERENCE GROUPS

($N = 52$)

Source	*df*	Mean Square	F	$p<$
Anxiety	1	40.69	8.71	.01
Preference	1	0.08	0.02	*ns*
Interaction	1	1.23	0.26	*ns*
Within Cells	48	4.67		

---

[7] The most deviant break, namely, that of 45%–55%, occurred for the variable of absolute preference level.

TABLE 5

Mean Levels of Arousal to Music for Social Extraversion and Preference Groups
($N = 11$ per cell)

	High Social Extraversion	High Social Introversion
High Preference	3.45	5.27
Low Preference	4.27	3.91

TABLE 7

Mean Levels of Arousal to Music for Social Extraversion and Anxiety Groups
($N = 15$ per cell)

	High Social Extraversion	High Social Introversion
High Anxiety	3.33	6.13
Low Anxiety	3.93	2.67

separated in terms of high vs. low level of absolute preference for the jazz. The same kind of analysis is shown in Tables 5 and 6 for social extraverts vs. social introverts. The preference level scores in each case were built as follows. An individual's absolute preference ratings for each piece on the seven-point scale were added to form a single combined estimate of preference for the set of musical selections as a whole, high numerical scores representing a high degree of liking for the jazz. The arousal level scores, in turn, consisted of an individual's sum score for all three selections combined. This is the case wherever arousal level scores are presented, unless otherwise specified. Our conclusion from Tables 3 to 6 is that preference level does not, by and large, influence arousal effects. In both cases, the preference variable alone is clearly without significance. The interaction, in turn, while clearly insignificant in the case of anxiety level and preference, shows a slight trend, although still being insignificant, in the case of social extraversion and preference. Since no clear preference effects were demonstrable, no reason remained for separating our *S*s on the basis of preference level.

Turning to the effects of personality, we find a strong relationship between level of manifest anxiety and symbolic sexual arousal (Tables 3 and 4), those *S*s who declare themselves more anxious being more arousable to the music. No comparable relationship occurs when one considers social extraversion alone (Tables 5 and 6). Such a finding is irrelevant to our predictions, however, since the hypotheses to be tested concerned interactive effects between social extraversion and anxiety level, and made no predictions concerning social extraversion per se. The necessary method, then, given the absence of preference effects, was to consider the joint effects of anxiety level and social extraversion–introversion on arousal.

As we see in Tables 7 and 8 these effects are considerable. Again, anxiety level taken by itself is positively related to arousal. But most noteworthy is the very strong

TABLE 6

Analysis of Variance, Arousal to Music for Social Extraversion and Preference Groups
($N = 44$)

Source	*df*	Mean Square	*F*	$p<$
Social Extraversion	1	5.82	1.08	*ns*
Preference	1	0.82	0.15	*ns*
Interaction	1	13.09	2.42	*ns*
Within Cells	40	5.40		

TABLE 8

Analysis of Variance, Arousal to Music for Social Extraversion and Anxiety Groups
($N = 60$)

Source	*df*	Mean Square	*F*	$p<$
Social Extraversion	1	8.81	2.23	*ns*
Anxiety	1	30.81	7.80	.01
Interaction	1	62.03	15.70	.001
Within Cells	56	3.95		

interaction (significant beyond the .001 level) between level of manifest anxiety and social extraversion–introversion in determining arousal. Social introverts high in manifest anxiety level exhibit a *greater* degree of symbolic sexual arousal than social extraverts whose anxiety level is the same. On the other hand, social introverts low in manifest anxiety level show a *lesser* degree of such arousal than do social extraverts of the same low anxiety level. Social extraversion–introversion per se had no effect on arousal, then, because its relationship to arousal strongly depends on the *S*'s level of manifest anxiety. For while those social introverts high in anxiety level show the greatest arousal to the music, social introverts low in level of anxiety show the least arousal, with both high and low anxiety social extraverts falling between these two extremes. These findings confirm the two hypotheses of the investigation.

Examples of Stories to Music

To provide a picture of the kinds of stories told by members of these four groups, here are typical stories by individuals of each type. The stories are all in response to the third musical selection—the piece intermediate in over-all arousal level.

High anxiety social introvert stories. Somebody is stirring a huge pot of boiling molasses which thickly swells and plops, sending off rich fumes which fill the room and hypnotize the stirrer. He becomes more and more drugged and entranced as he stirs, until finally he climbs into the pot with the molasses and lets himself be surged around and down into it.

A girl is being chased down an empty street by four men. She runs wildly into the apartment house in which she lives and takes the elevator to the fourth floor. She runs to her room and locks the door herself in. In a few minutes there is banging on the door. The men's banging shakes it and in desperation she shrieks for help. Just then a man from one of the nearby apartments steps out of his doorway and upon seeing him the other men flee.

There is a man with a chain in his hand, approaching a horse. At first, the horse is frightened and shys away. The man talks calmly to the horse, but it does no good. He slaps the horse with the chain. The horse rears up, kicking, whinnying. The man hits him harder. Finally the man realizes that he can't beat the horse so he goes back to soft-spoken words. This method works and the man and horse are confident again. The man tries to mount but the horse throws him off. Eventually, after a long struggle the horse gives in to the man's weight and they are friends.

Beautiful woman in green satin dress, half intoxicated, half exhausted from lack of sleep, lackadaisically starts climbing flight of stairs to her room. Loses balance and falls—dead! Her husband returns and sees woman in purple dress—doesn't recognize her. Steps over dead body; climbs stairs, and goes to bed. He never wakes up.

Low anxiety social introvert stories. This is the story of a young man who loved music. He was not from a wealthy or a poor family. He came from a small northern town of a middle-class family. Of course his parents were against him going into the music profession. His father had been with a well-known band in the late 1920's. He had always wanted to go on to play with his own band. He never made it, and consequently he was afraid that his son would find himself in the same position.

A college student has the "blues." The past week has been a hard one for him. He flunked the exam he had, his roommate got sick, and had to be sent home, he couldn't get a date for the big dance tonight. It's a lovely Saturday evening as he sits in his room trying to study while all his friends are at the dance.

A new family moved into the rough section of town. They were poor but respectable. The mother watched with horror her son grow into a rude, troublesome teenager. His talk, when he spoke to her, was of playing "chicken," rock and roll, and the gang. She helplessly saw him take up smoking. She was sure he drank. His friends were tough; dressed in leather jackets with gaudy symbols on the back. Every time she heard the scream of brakes, her thoughts were of her son. Gang wars were the curse of the neighborhood. Surely, she thought, my son could never hurt anyone. I've tried to teach him what's right.

To play such music in such a way as to make it speak, not only does the musician have to have talent and the right instrument but also some experience to remember or feel right at the moment. The player has lost a loved one, be it a girl friend or his mother, etc. His sadness is told to the audience or perhaps one lonely member of it. The audience feels the beauty of his music, the lonely man hears it in the distance, but neither understand the powerful emotion behind a player of the blues.

High anxiety social extravert stories. This is the nonconformist who is ruled by his passions and emotions. He won't listen or be put into the mold made for him. His only outlet is to express the freedom of feeling—to cleanse his system of the agitation and bitterness within. He cares for no one it seems and so feels driven to show this subjective rather than objective view. He is agitated by things around him and this freedom of expression is a way of revolt for him. He has something new to say and says it regardless of public opinion and criticism. His thoughts are vivid yet rash, but isn't that so of anyone who has had a new idea and wishes to convey it? He must suffer for his outwardness.

Ethel welcomed her husband home with a smile and words of kindness. She seductively led him to the love seat, begged him to make himself comfortable, and then announced that she had a few things to say. Poor Bill—he wasn't expecting the torrent of words that rained at him—they were more like a continuous blast of thunder. Higher and higher, faster and faster went Ethel's voice as she accused him of everything from infidelity to acts of treason against her flower garden. Bill couldn't say a word—not only could he not talk louder or faster, but there was no opportunity for him to even begin. Finally she stopped and stood staring at him, laughing uncontrollably.

It's almost as if the trumpet can talk—the player is telling the world they're not the only ones who have troubles—he's blaring his aloud and baring his heart to all—some get saved (increased tempo and other instruments) but others keep dragging on—the color blue is suggested for it depicts the "blues" and this person has definitely got them! He's reassuring though—for one realizes everyone is in the same boat.

I picture a man, hurt yet unable to strike out in retaliation. Instead he roams the street, drunk, poor, another bum who wishes there was some way out, but doesn't have the strength to look for it.

Low anxiety social extravert stories. In order to get to the telephone building in Bowdoin Square we had to walk through Scully Square. As we were passing an open door Greg suddenly pulled my hand so hard that I had no choice but to follow him. Inside a very dimly lit room there was a jukebox playing, and the men clustered about the bar were swaying with the rhythm. I tried to pull Greg away from the door, but he was fascinated, more by the people than the music. Suddenly the sound sharpened and I tugged Greg away but we walked away quite slowly so we would hear the music.

It is a dark, humid, heavy evening in early August when everything seems to be sticky and one feels that he just can't do anything. But one is just wasting time. Restlessness and boredom—where to go and what to do. One rises from the chair in which one was just slouching and gets a good book—can't concentrate, no interest. Go for a ride—stop get a cool drink—visit some friends to say hello—return home and the same sluggishness engulfs the mind and body. What to do and where to go—just to release the nothingness that the stillness and the calm that the whole day produced.

The large trees withheld the force of the cruel wind that blew through the woods. Baby Bunny, who should have known better, had wandered away from home. So interested was he in the little things he found along the ground that he hadn't heard the baying of the hunting hounds. All of a sudden, he hears a loud bark, and looking up he just stood still—not knowing what to do. The hound was also surprised to see such a little fellow out alone on such a cold day—barked again—Baby Bunny at that point decided to run—and away he went. The hound thought he would have a little fun—so chased. . . .

It was Sunday and all the workers of the farm had the day off. Today was a special day for it was the annual picnic day. Everyone dressed casually and came down along the river banks. There they all stretched out, opened their picnic baskets, and ate as they watched the boats go by them. The small boats went along at their leisure, in no particular hurry to reach their destination. A few larger ones came by, making more noise and causing more excitement among the spectators. As the day drew to a close a group of happy people gathered together their belongings to prepare for the return to the farm.

Control for Possible Differences in Basal Level of Arousal

Are the strong arousal effects for the high anxiety social introverts truly a function of listening to the music? Might they not simply result from a higher basal level of symbolic sexual arousal for these *S*s? In such a case, their verbal output would reflect this high arousal level, whether the stimulus be high or low arousal music. Several further analyses, however, suggest that listening to the music is in fact influencing this group's arousal level, although they also seem to have a somewhat higher basal level than the others.

First, recall our finding, for the sample as a whole, that over-all arousal was lowest for Piece A, highest for B, and intermediate for C. A similar analysis was done for the 15 high anxiety social introverts as a separate group, yielding the same results: mean arousal levels for these *S*s being 1.33 for A, 2.45 for B, and 2.33 for C. Once again, statistical evaluation was made in terms of the difference between each *S*'s arousal scores for any two musical pieces, using a t test to determine whether to reject the null hypothesis of zero difference. For B–A this t value was 2.45 and for C–A it was 2.20, both being significant beyond the .05 level. For C–B the t value was zero, since the mean arousal difference score for C–B likewise was zero. Thus, the high anxiety introverts, like the sample as a

TABLE 9

MEAN LEVELS OF AROUSAL TO SELECTION A FOR SOCIAL EXTRAVERSION AND ANXIETY GROUPS
(N = 15 per cell)

	High Social Extraversion	High Social Introversion
High Anxiety	0.73	1.33
Low Anxiety	0.66	0.66

whole, show greater arousal to selections B and C than to A.

Second, and more important, recall that our analysis of variance demonstrating an interaction between anxiety level and social extraversion–introversion in determining arousal level used an arousal score based on all three musical selections combined. Suppose now we were to carry out three successive analyses of this kind, using arousal scores for each selection alone as the dependent variable. If a higher basal level of arousal is the only cause of the high score for anxious social introverts, then the interaction effect should be significant for the first piece—i.e., the one which is lowest in arousal level for the group as a whole—as well as for the others. If, on the other hand, the significant interaction for all three pieces combined is found to depend on the second and third pieces but not on the first piece, this will indicate that the change from low to high arousal music caused a greater increase in arousal for the anxious introverts than for the other groups, that, in other words, the music had an effect.

TABLE 10

ANALYSIS OF VARIANCE, AROUSAL TO SELECTION A FOR SOCIAL EXTRAVERSION AND ANXIETY GROUPS
(N = 60)

Source	*df*	Mean Square	*F*	$p<$
Social Extraversion	1	1.35	1.32	*ns*
Anxiety	1	2.00	1.96	*ns*
Interaction	1	1.37	1.34	*ns*
Within Cells	56	1.02		

TABLE 11

MEAN LEVELS OF AROUSAL TO SELECTION B FOR SOCIAL EXTRAVERSION AND ANXIETY GROUPS
(N = 15 per cell)

	High Social Extraversion	High Social Introversion
High Anxiety	1.60	2.40
Low Anxiety	1.66	1.40

Tables 9 through 14 present the relevant information. We find that the anxious social introverts' degree of arousal in stories to Piece A (Tables 9 and 10) is not significantly different from that of the other three groups: the interaction F is 1.34. For Piece B (Tables 11 and 12), on the other hand, this interaction F approaches significance, being 2.85 (passes the .10 level of significance); and for Piece C (Tables 13 and 14), this interaction F is highly signifi-

TABLE 12

ANALYSIS OF VARIANCE, AROUSAL TO SELECTION B FOR SOCIAL EXTRAVERSION AND ANXIETY GROUPS
(N = 60)

Source	*df*	Mean Square	*F*	$p<$
Social Extraversion	1	1.06	0.71	*ns*
Anxiety	1	3.26	2.17	*ns*
Interaction	1	4.28	2.85	.10
Within Cells	56	1.50		

TABLE 13

MEAN LEVELS OF AROUSAL TO SELECTION C FOR SOCIAL EXTRAVERSION AND ANXIETY GROUPS
(N = 15 per cell)

	High Social Extraversion	High Social Introversion
High Anxiety	0.93	2.40
Low Anxiety	1.53	0.60

TABLE 14

Analysis of Variance, Arousal to Selection C for Social Extraversion and Anxiety Groups ($N = 60$)

Source	*df*	Mean Square	*F*	*p*<
Social Extraversion	1	1.06	0.66	*ns*
Anxiety	1	5.40	3.37	.08
Interaction	1	21.60	13.50	.001
Within Cells	56	1.60		

cant ($F =$ 13.50, which passes the .001 level). Thus, the greatest tendency for our anxious social introvert group to exceed the others in arousal occurs for Piece C—the piece which, we recall, is intermediate in arousal value for the sample as a whole. The results for Selection B—the piece which is highest in overall arousal value—show a trend in the same direction. Finally, no such trend occurs in the case of the music lowest in arousal value, Piece A.

The difference in results for Pieces B and C becomes understandable from a consideration of the cell means in Tables 11 and 13, respectively. In both cases, the arousal level for high anxiety social introverts is 2.40, and that for low anxiety social extraverts is in the neighborhood of 1.50. The fact of Piece B's being highest in general arousal value pushes up the mean arousal for the other two cells of Table 11, whereas these cells remain low for Piece C—the selection intermediate in overall arousal value (Table 13). The stimulus value of the music with regard to arousal level, then, has its greatest influence for the pieces that are lowest and highest in arousal value—Selections A and B, respectively. Personality has its greatest chance to influence arousal when, stimulus-wise, the music has an intermediate degree of arousal value: namely, in the case of Piece C. Here, the groups whose arousal reactions to the music are strong because of personality factors remain strong. Those cells, on the other hand, where personality would lead to weak arousal reactions, show a drop.

TABLE 15

Mean Number of Words in Stories to Music for Social Extraversion and Anxiety Groups ($N = 15$ per cell)

	High Social Extraversion	High Social Introversion
High Anxiety	211.33	224.00
Low Anxiety	212.67	182.67

In sum, the heightened arousal for anxious social introverts seems indeed to be an effect of listening to the music. It does not significantly occur for the music lowest in arousal value, but appears for the other selections.

Control for Possible Differences in Verbal Facility

One might argue that the greater arousal for high anxiety social introverts is not due to group differences in the tendency to use sexual symbolism in one's stories, but rather simply to differences in story length. That is, the chances might be greater for sexual symbolism to appear in the stories of anxious social introverts simply because they write longer stories. To evaluate this point, the number of words in all three stories combined was calculated for the members of the four groups under discussion. Table 15 presents means for these groups and Table 16 an analysis of variance, indicating that no significant varia-

TABLE 16

Analysis of Variance, Number of Words in Stories to Music for Social Extraversion and Anxiety Groups ($N = 60$)

Source	*df*	Mean Square	*F*	*p*<
Social Extraversion	1	11.26	0.33	*ns*
Anxiety	1	59.73	1.74	*ns*
Interaction	1	69.54	2.03	*ns*
Within Cells	56	34.31		

Note.—To facilitate computing the analysis of variance, individual scores were divided by 10.

tions in story length differentiate our groups. Comparable analyses of variance done for the length of first, second, and third stories separately likewise showed no significant variations in story length—no F value exceeding the largest one in Table 16. The stories written by high anxiety social introverts, then, were not significantly longer than those by members of the other three groups. Some tendency of this kind may be operating, however, since—as shown in Table 15—the means are in the expected direction, although not significantly so.

Agreeing Response Set

It should be noted that all of the Manifest Anxiety scale items (the Maudsley Neuroticism scale) are keyed in such a way that "yes" responses cause a high anxiety score, while two-thirds of the social extraversion items (the Maudsley Extraversion scale) are keyed in such a way that "yes" responses cause a high social extraversion score. If an agreeing response set (Cohn, 1956; Couch & Keniston, 1960) were operating in our sample, one hence would expect a significant positive correlation between the two scales, since a person scoring high in manifest anxiety would score in the middle on social extraversion. But the correlation between scores on the Manifest Anxiety and Social Extraversion scales was −.15, which is nonsignificant for the sample of 88 for whom scores were available, and even negative in direction.

A more important indication that agreeing response set doesn't exert a potent influence on arousal is the complicated relationship between arousal, social extraversion–introversion, and manifest anxiety. The most "yes" responses are given by high anxiety extraverts, and yet they fall in the middle on the arousal distribution, being excelled by *S*s (high anxiety introverts) who give more "no" responses.

These points argue against the influence of a "yea-saying" response set on arousal.

Motivational Differences Other Than in Symbolic Sexual Arousal

It is clear from the illustrative stories presented above that aspects of motivational content other than symbolic sexual arousal also may differentiate our personality groups. The low anxiety social introverts, for example, would seem to be deniers or repressors of affective involvement. Their stories strive to present as literal, descriptive renderings of the music as is possible. Further, the high anxiety social extraverts seem to be individuals whose contacts with other people are covertly dominated by the hostile wish to hurt them—as expressed in story fantasy. It is possible that while the main motivational effect of music for high anxiety social introverts is to elicit covert sexual expression, music's main motivational influence for high anxiety social extraverts is to stimulate covert expression of hostility. In each case, the music would permit displaced, implicit expression of a kind of motive which is denied outlet at the overt level: the anxious social introvert inhibiting sexual needs for others, the anxious social extravert inhibiting aggressive feelings toward others and counterphobically insisting that he likes people.

This points toward the idea that verbal expression of social extraversion, if accompanied by high manifest anxiety, may be just as indicative of the failure of a mode of social adaptation as the accompanying of social introversion by high anxiety. The anxious social extravert, rather than wanting to be with other people, really may hate them! Further exploration of this possibility must wait upon continuing research.

Emotional Extraversion–Introversion

The correlation between social extraversion and emotional extraversion scores was .16, which is nonsignificant for our sample of 88. These two scales do in fact seem to be measuring different things. A borderline correlation, in turn, exists between the Emotional Extraversion and Manifest Anxiety scales: the r of .23 passing the .05 level with 88 *S*s. As one well might expect, the tendency to be generally expressive of one's feelings (emotional extraversion) has some relationship with the tendency to express feelings of worry and depression in particular (manifest anxiety).

TABLE 17

MEAN LEVELS OF AROUSAL TO MUSIC FOR EMOTIONAL EXTRAVERSION AND PREFERENCE GROUPS

(N = 11 per cell)

	High Emotional Extraversion	High Emotional Introversion
High Preference	4.36	3.27
Low Preference	5.18	3.27

This last correlation indicates a need to investigate whether the emotional extraversion dimension behaves at all like manifest anxiety level in its relationships to symbolic sexual arousal. The relevant information is presented in Tables 17 through 22. Contrasting emotional extraverts and introverts with high or low general preference for the jazz selections, we find from Tables 17 and 18 that the same general effect occurs as in the case of manifest anxiety, although with less intensity. Again, preference level makes no difference, and emotional extraverts tend to be more aroused to the music than emotional introverts. When one pits anxiety level and emotional extraversion against each other, however (Tables 19 and 20), we find a clearly significant effect occurring only for the former. Finally, contrasting emotional and social extraversion (Tables 21 and 22), we find a trend toward significance for the interaction effect ($p < .07$) : those who are emotional extraverts and social introverts being highest in arousal level. These are individuals who claim that they are emotionally expressive but yet have declared that they keep apart from other people. We recall that this same interaction is highly significant when anxiety level and social extraversion–introversion are contrasted.

TABLE 18

ANALYSIS OF VARIANCE, AROUSAL TO MUSIC FOR EMOTIONAL EXTRAVERSION AND PREFERENCE GROUPS

(N = 44)

Source	*df*	Mean Square	*F*	$p<$
Emotional Extraversion	1	24.75	5.04	.05
Preference	1	1.84	0.37	*ns*
Interaction	1	1.84	0.37	*ns*
Within Cells	40	4.91		

TABLE 19

MEAN LEVELS OF AROUSAL TO MUSIC FOR EMOTIONAL EXTRAVERSION AND ANXIETY GROUPS

(N = 13 per cell)

	High Emotional Extraversion	High Emotional Introversion
High Anxiety	5.92	4.23
Low Anxiety	3.46	3.23

These findings suggest that manifest anxiety level and emotional extraversion–introversion have something in common—namely, the tendency to express one's various feeling states. But the expression of a particular kind of feeling state—namely, anxiety, worry, and depression—seems most directly related to symbolic sexual arousal, since we obtain much stronger arousal effects when varying anxiety level than when varying emotional extraversion. This is consistent with what our theory leads us to expect, since a defense's success or failure was defined in terms of its relation to a high or low level of anxiety in

TABLE 20

ANALYSIS OF VARIANCE, AROUSAL TO MUSIC FOR EMOTIONAL EXTRAVERSION AND ANXIETY GROUPS

(N = 52)

Source	*df*	Mean Square	*F*	$p<$
Emotional Extraversion	1	12.02	2.71	*ns*
Anxiety	1	38.94	8.79	.01
Interaction	1	6.94	1.57	*ns*
Within Cells	48	4.43		

TABLE 21

Mean Levels of Arousal to Music for Social Extraversion and Emotional Extraversion Groups

($N = 17$ per cell)

	High Social Extraversion	High Social Introversion
High Emotional Extraversion	3.53	5.41
High Emotional Introversion	3.71	3.59

particular. And failure or success of the introversion defense, in turn, was the crucial determinant of strength of symbolic sexual arousal.

Sex Differences in Arousal

Finally, our results seem to shed light on Beardslee and Fogelson's (1958) finding that female undergraduates tend to be higher in symbolic sexual arousal to music than do male college students. If the strongest determinant of such arousal is fantasy displacement of sexual needs that are hindered from more direct expression, then the contrasting social norms in our culture regarding the sexual activities of unmarried men and women would lead us to expect the result reported. The stricter societal prohibitions against girls for direct sexual expression should lead to a stronger need in the case of unmarried women than men for displacement arousal in response to music.

TABLE 22

Analysis of Variance, Arousal to Music for Social Extraversion and Emotional Extraversion Groups

($N = 68$)

Source	*df*	Mean Square	*F*	$p <$
Social Extraversion	1	13.24	2.81	*ns*
Emotional Extraversion	1	11.53	2.44	*ns*
Interaction	1	17.00	3.60	.07
Within Cells	64	4.72		

Summary and Conclusions

A study was carried out to test the following theory about music and personality: (*a*) One function of listening to music is to permit the symbolic expression of sexual impulses. (*b*) The strongest impetus toward such expression arises for persons who claim to be socially introverted, but whose attitude of introversion does not constitute a successful defense as indicated by the presence of a high level of anxiety. Since the need for others remains and yet its direct sexual expression has been prohibited, the only open path is that of fantasy displacement. (*c*) A weaker impetus toward such expression arises for those depicting themselves as socially extraverted, simply as a reflection of their social involvement. Since direct sexual expression is still an available channel, however, symbolic expression is not the only permitted route and hence should be less intense than in the first case. (*d*) The weakest impetus toward such expression exists for those whose attitude of introversion is a successful mode of defense as indicated by the absence of anxiety, since such success means their need for other persons really has been overcome. Hence, a minimum of symbolic expression is required.

Two hypotheses were derived from this theory: (*a*) Among those high in admitted anxiety level, *S*s describing themselves as socially introverted will show a greater degree of symbolic sexual arousal to music than those calling themselves socially extraverted. (*b*) Among those low in admitted anxiety level, *S*s calling themselves socially introverted will show a lesser degree of symbolic sexual arousal to music than those considering themselves socially extraverted.

To evaluate these hypotheses, an experiment was devised involving assessment of symbolic sexual arousal from stories inspired by listening to music. The results, using female undergraduates as *S*s, strongly confirmed both predictions. A number of points of interpretation were treated in the course of describing the procedures and results, including the following: the assumed relevance of the story-scoring categories to symbolic sexual arousal was considered. Strong relations were found between the independently judged sexual stimulus values of the musical selections and their respective symbolic sexual arousal potentials as inferred from story content. The possibility was examined of accounting for the results as due to differences in basal level of arousal rather than in the effects of the music, and it was concluded that the music itself was the influential factor. In addition, the obtained results were not found to be due to group differences in verbal facility, nor could they be the product of an agreeing response set. Comments were made on motivational differences other than in symbolic sexual arousal. Finally, relations to emotional extraversion and an interpretation of sex differences in symbolic sexual arousability also were discussed.

The results of the present research seem to suggest the fruitfulness of further work on music's motivational functions for the individual.

REFERENCES

BEARDSLEE, D. C., & FOGELSON, R. Sex differences in sexual imagery aroused by musical stimulation. In J. W. Atkinson (Ed.), *Motives in fantasy, action and society.* New York: Van Nostrand, 1958. Pp. 132–142.

CLARK, R. A. The effects of sexual motivation on phantasy. In D. C. McClelland (Ed.), *Studies in motivation.* New York: Appleton-Century, 1955. Pp. 44–57.

COHN, T. S. Relation of the F scale to a response set to answer positively. *J. soc. Psychol.,* 1956, **44,** 129–133.

COUCH, A., & KENISTON, K. Yeasayers and naysayers: Agreeing response set as a personality variable. *J. abnorm. soc. Psychol.,* 1960, **60,** 151–174.

ERIKSON, E. H. *Young man Luther.* New York: Norton, 1958.

EVANS, CATHERINE, & MCCONNELL, T. R. *Minnesota T-S-E inventory.* (Rev. ed.) Princeton: Educational Testing Service, 1957.

EYSENCK, H. J. The questionnaire measurement of neuroticism and extraversion. *Riv. Psicol.,* 1956, **54,** 113–140.

FARNSWORTH, P. R. *The social psychology of music.* New York: Dryden, 1958.

FENICHEL, O. *The psychoanalytic theory of neurosis.* New York: Norton, 1945.

FREUD, S. Three contributions to the theory of sex. III. The transformations of puberty. In A. A. Brill (Ed.), *The basic writings of Sigmund Freud.* New York: Modern Library, 1938. Pp. 604–629. (a)

FREUD, S. The interpretation of dreams. In A. A. Brill (Ed.), *The basic writings of Sigmund Freud.* New York: Modern Library, 1938. Pp. 179–549. (b)

GUTHEIL, E. A. *The handbook of dream analysis.* New York: Liveright, 1951.

JENSEN, A. R. The Maudsley personality inventory. *Acta psychol.,* 1958, **14,** 314–325.

MCNEMAR, Q. *Psychological statistics.* (Rev. ed.) New York: Wiley, 1955.

SNEDECOR, G. W. *Statistical methods.* Ames, Iowa: Iowa State Coll. Press, 1946.

SOPCHAK, A. L. Individual differences in responses to different types of music, in relation to sex, mood, and other variables. *Psychol. Monogr.,* 1955, **69** (11; Whole No. 396).

SOPCHAK, A. L. Relation between MMPI scores and musical projective test scores. *J. clin. Psychol.,* 1957, **13,** 165–168.

SULLIVAN, J. W. N. *Beethoven: His spiritual development.* New York: Mentor, 1949.

TAYLOR, JANET A. A personality scale of manifest anxiety. *J. abnorm. soc. Psychol.,* 1953, **48,** 285–290.

WALLACH, M. A. Art, science, and representation: Toward an experimental psychology of aesthetics. *J. Aesthet.,* 1959, **18,** 159–173.

(Received April 18, 1959)

Vol. 74, No. 8 Whole No. 495, 1960

Psychological Monographs: General and Applied

OPPOSITES STRUCTURES, DEFENSES, AND ATTITUDES[1]

HELEN PEAK, BARBARA MUNEY, AND MARGARET CLAY
University of Michigan

THIS is one of a series of studies in which various relations between psychological events have been examined as determinants of attitude and attitude change. The opposites relation is of interest for a number of reasons. In the first place, scattered bits of evidence (Woodworth, 1938; p. 344; Riess, 1946; Kreezer & Dallenbach, 1929) suggest that the opposites pattern may emerge as a result of the differentiation of a similarity dimension and consequently may involve greater distances between parts of the structure than do less differentiated similars. The nature of generalization within an opposites structure has been explored by one of us (Peak, 1958b, 1958c).

The opposites concept has also been invoked in discussions of personality characteristics, most particularly as an aspect of the authoritarian syndrome which has been said to involve extreme tendencies to dichotomize (Adorno, Frenkel–Brunswick, Levinson, & Sanford, 1950). Again, the presence of an opposites structure is implicit in the descriptions of certain defense mechanisms. For example, projection, defined as the attribution of characteristics or feelings possessed by an individual to persons and objects in the outside world, probably requires a dichotomization of the concepts of the self and certain others, for it should be easier to blame others with impunity if they are highly differentiated from the self. Reaction formation or reversal, involving the overt expression of an act or feeling opposite the one being defended against, also required examination from this point of view.

A pretest and an earlier study (Peak 1959a) indicated that certain measures assumed to reflect opposites structure tended to be positively correlated. The pretest also supported the notion that a preference for projection was positively related to opposites measures but that reversal or reaction formation preference, contrary to expectations, had a negative relation to these measures. The investigation to be reported sought additional evidence of a general tendency to dichotomize and examined the relations of opposites measures to certain defense preferences, to ingroup, outgroup, and self-attitudes, to scores on the California F Scale, and to other variables, with the expectation that methods of dealing with psychological distance might prove to be a common determinant of many such psychological processes.

Definition of Opposition

Similarity has usually been defined as a relationship in which events share one or more common properties. Opposed events, like similars, also have common properties, the principal difference residing in the greater psychological distance between opposites, a distance which has become sufficiently large to produce subcategories within the common dimension. But it should be noted that such events are opposites only as long as the emerging subcategories retain positions on the same dimension. Opposites must be alike in some degree. For example, clean and dirty may in some cases be seen as possessing different qualitative properties (subcategories), though from another point of view they are simply different amounts of the same property. Again, liking and disliking share the property of affect which appears to be structured as a bipolar continuum passing through a zero point representing no affect or evaluation. Presumably, any continuum made up of degrees of some property may under certain conditions become sufficiently differentiated so that its ends come to constitute subcategories and thus an opposites structure, provided some relation remains between the dichotomies.[2]

---

[1] This investigation was carried out under a Contract between the Office of Naval Research and the University of Michigan, Project Nonr 1224 (10) NR 171-039. The senior author was in residence at the Center for Advanced Study in the Behavioral Sciences when this was written.

Defenses

In the following pages frequent reference will be made to the defenses of projection and reaction formation or reversal. These concepts have been assigned many different meanings in the literature. In this discussion they will be used to refer to processes of the kind described on p. 1 and identified by the Blum Defense Preference Inventory. The extent to which the results obtained here are generalizable to other measures of projection and reversal can be determined only by further observations.

Hypotheses

The following hypotheses were tested:

I. Different measures of opposites structure will tend to be positively related. Since no psychological characteristics are ever completely general over all situations, it was not expected that all such correlations would be significant. It was rather supposed that the correlations would be higher as the methods of measurement and the nature of the situations involved in the indices became more similar. For example, there should be a higher correlation between two indices based on deviations from the midpoint of a Semantic Differential scale than between two derived from such a scale in one case and from opposites associations in the other. Furthermore, similarity in content of the indices should also be related to the size of the correlation between them. However, in the absence of pre-experimental commitments about the degree of similarity of each index to the others, no predictions were made about the relative amounts of correlation between specific measures here employed.

II. Preference for projection will be negatively correlated with preference for reversal or reaction formation. This relation has been observed in an earlier study by Goldstein (1952) and more recently in our pretest.

IIIa. Preference for projection will be positively related to a tendency to opposites structuring.

IIIb. Preference for reversal will be negatively related to opposites structuring.

We have already noted the rationale for Hypothesis IIIa; the prediction of IIIb was made on the basis of pretest results without a clear understanding of why this should be so. This will be discussed later.

IV. Those who prefer projection will have less favorable attitudes toward the outgroup and more favorable attitudes toward the ingroup than those who do not prefer projection; i.e., there will be a negative correlation between preference for projection and outgroup attitude and a positive correlation between ingroup attitude and projection. Furthermore, projection preference and difference in ingroup and outgroup attitudes will be positively correlated.

The greater negativity toward the outgroup and more favorable attitude toward the ingroup would follow from the familiar assumption that a person's hostility is displaced from the ingroup to the outgroup, which is at once hated and blamed for hating. The blame is said to be a rationalization to the effect that if outgroup members are hostile it is justifiable to return the hostility. Since the distance between ingroup and outgroup attitudes may be thought of as a measure of the tendency to dichotomize these concepts, the greater ingroup–outgroup difference predicted for the projector would be expected, for it is easier to express hostility to the outgroup with impunity if the outgroup and ingroups are clearly separate. It will be suggested later that the increased distance between ingroup and outgroup for projectors, as well as the reversal results, could be explained in terms of a style of dealing with *disparity* or distance which is the basis of a motive system.

---

[2] These relations are discussed in detail in another paper (Peak, 1958a), including the fact that under certain conditions the reactions to the terms of an opposites structure may cancel each other.

V. Those who prefer reversal will have more favorable outgroup attitudes than those who do not prefer it; there will be a positive correlation between outgroup attitudes and reversal preference.

Since this defense is defined as involving responses in a direction opposite to those which are being defended against, it was expected that positive attitudes toward the outgroup would accompany high reaction formation preference. Positive attitudes are predicted as defensive reactions because negative attitudes toward outgroups are assumed to be more likely to require defensiveness and reversal.

VI. Women will tend to prefer reversal as a defense more often than will men and men will prefer projection more often than will women. This tendency was observed in the pretest and the suggestion was made that such a difference might result from the greater social penalties attached to the overt aggressiveness that tends to accompany projection of hostility, while the more positive responses resulting from reaction formation defenses in a hostility situation would be the more approved pattern of feminine behavior.

PROCEDURE

Undergraduate students at the University of Michigan served as subjects(*S*s). The data were collected during two one-hour sessions approximately one week apart. Men and women were tested at different times because of the difference in standard procedures on the Blacky Test for the two sexes. During the first session these tasks were performed: The Blacky Test, including the Defense Preference Inventory, the Kent–Rosanoff free association test, the California F Scale, and filling in a face sheet. At the second session *S*s continued with the free association test. They sorted 108 statements about Negroes, judging the degree of favorableness which the statements represented. And finally they rated 24 concepts on Semantic Differential scales.

Blacky Test

Standard procedures were followed in administering Blum's Blacky Test[3] and Defense Preference Inventory (Blum, 1950, 1956, 1957; Blum & Hunt, 1952). *S*s wrote a story about each picture after the projection of that picture on the screen. All stories about the pictures were completed before the Defense Preference Inventory was given. This involved exposing each picture a second time, and asking *S*s to rank five statements in the order of their appropriateness to the picture that had just been shown. Each of these standardized statements was intended to reflect a preference for either projection, reaction formation, avoidance, regression, or intellectualization. The statements referring to a given picture were ranked before the next picture was presented. A man administered the test to the men students; a woman, to the women students.

Conflict scores were derived from the stories which were coded by Blum's procedures. This coded material also provided data for an index of hostility. This was a count of the number of hostility items in responses to all the pictures.[4] Hostility scores were also expressed as a percentage of the total "ideas" written in the stories. Rules were made for classifying all the items in the stories into unit ideas. The coder had no information about the hostility or conflict scores of the *S* whose protocol was being coded, nor was she familiar with the Blacky scoring method. The number of these idea units in all the stories was counted and divided by the number of stories.

It was planned at the outset that two DPI (Defense Preference Inventory) scores would be used in analyzing the data. (*a*) An average rank of all the statements reflecting any one class of defense in response to all pictures. Thus, the reversal score was based on the average rank of the reversal statements, the projection score on the average of all projection ranks, etc. Because of an error in the order of DPI statements for two of the pictures (X and XI) in one section of women students, *total* DPI scores were derived from nine of the eleven pictures. (*b*) An average of the ranks of each class of statement in response to those pictures which Blum identified as hostility situations.[5] These hostility pictures were selected because it was thought likely that defense preferences based on hostility situations would have a clearer relation to ingroup–outgroup attitudes than would responses to the other pictures. This proved to be the case, although the direction of results was the same for both DPI indices. The major part of the analysis reported at this time is based on DPI scores derived from the hostility pictures, though reference will be made to certain relations involving DPI scores based on high conflict pictures. This is the third method of dealing with Blacky and one which Blum (1954) has employed.[6]

---

[3] We are grateful to Professor G. S. Blum for his advice on the use of the Blacky Test and to Mrs. Blum for training one of us to score the projective stories on which hostility scores and conflict scores were based.

[4] Blum's General Hostility Pattern (1957, p. 12) was modified for this purpose by Barbara Muney and the hostility score included items from the stories which fell into these classes: overtly expressed oral hostility, hostile interaction with parents in an anal context, resentment against the parents for lack of attention in oedipal context, readiness to experience anger or to engage in aggressive behavior in response to discipline, overtly expressed hostility, feelings of rejection in comparison with sibling, attraction to the sibling rivalry situation, and certain denial statements implying hostility.

Association Test: 1

One hundred words were projected onto a screen one at a time, and *S*s were instructed to write down on a sheet of paper the first word that came to mind. Approximately four seconds intervened between the exposure of each word. The list on this day consisted of the first 75 Kent–Rosanoff words plus 25 other words which were the same ones rated on Semantic Differential scales (see below).

California F Scale

Thirty items were administered in the usual manner, but the reported results are based on 21 ego-defensive items (Katz, McClintock, & Sarnoff, 1957), leaving out items placed by the authors of the scale (Adorno et al, 1950) in categories of "authoritarian submission" and "conventionalism." In those cases where both forms of the F scale were correlated with the same variables only small differences were obtained.[7] High scores mean high F tendencies.

Association Test: 2

A second free association test, administered at the beginning of the second session, included the last 25 words of the Kent–Rosanoff list which had not been given the first day.

The association responses were scored as opposites, complements, and others. A score based on the total number of opposites associations to the Kent–Rosanoff list and to 25 additional words is reported.

Sorting Statements about Negroes

One hundred and eight statements about Negroes were sorted into nine piles on the basis of judgments of the extent to which the statements were favorable or unfavorable toward the Negro. The Hovland and Sherif (1952) procedure was followed except that 9 instead of 11 categories were used. A *bimodality index* was determined by adding the number of items placed in categories 1+2+8+9, dividing the sum by 4 and subtracting the sum of statements in categories 3+4+5+6+7 divided by 5.

An analysis of the mean frequency of items sorted into each of the nine categories by *S*s with bimodality scores in the upper half of the distribution and by those in the lower half showed the expected bimodal pattern in the first group, i.e., more items in end categories than in the middle. The persons with low bimodality scores produced rectangular distributions.

Semantic Differential Ratings

The last task in this session involved rating 24 concepts on 12 scales (Osgood, 1955).[8] These concepts included a neutral filler word at the beginning, the six concepts from which ingroup and outgroup attitudes were derived (NEGRO ROOMMATE, NEGRO NEIGHBOR, FOREIGNER, FRIEND, FAMILY, AMERICANS), the ME concept plus the following: DANGER, SAFETY, CLEAN, DIRTY, HONESTY, DECEIT, MASTER, SERVANT, THRIFTY, WASTEFUL, SUCCESS, FAILURE, DOMINANT, SUBMISSIVE, ENEMY, EARLY. A random sequence of rating these concepts was used. Indices derived from these measures will be described with the results.

---

[5] These were Pictures II, III, IV, VII, and VIII.

[6] After completing the analysis with these two DPI indices, two further indices were devised in order to discover for future use whether there were more effective ways of using DPI measures. The first supplemental method involved averaging the DPI ranks for those pictures on which a given *S* had a conflict score equal to or greater than *two*. The second supplementary score was based on the three pictures having the highest correlations with a criterion measure, the difference in ingroup minus outgroup attitude scores.

Intercorrelations of all items are found in the ONR Technical Report No. 7 (Peak, 1959b). Both supplementary scores yielded results in the same direction as DPI indices (*a*) and (*b*), but they had generally somewhat less significant relations to other variables.

[7] Seven statements were added to the test near the beginning, phrased in such a way that disagreement would be expected of those with authoritarian tendencies. This was done in order to discourage a tendency to answer all questions simply by taking a set for or against the statements. These seven items were not scored.

[8] The scales were labelled with these descriptive terms: loud–soft, bad–good*, small–large #, wide–narrow #, awful–nice*, pungent–bland, pull–push, distasteful–tasty*, deep–shallow #, toward–away from, angular–round, fair–unfair*. Evaluative scores were based on scales marked with an asterisk; potency scores on scales marked with #.

Aptitude Measures

Scores on the 1949 edition of the American Council of Education Tests were available for most of this sample. The tests had been given by the Michigan Bureau of Psychological Services early in the academic career of most of the students. A large proportion of our sample was from the Sophomore class at the time of the observations and had taken the test the year before.

RESULTS

Aptitude as a Variable

The aptitude scores of subgroups resulting from dividing the sample simultaneously on projection and on reversal scores are found in Table 5. There were no reliable differences in the aptitude scores of the subgroups when men and women were treated separately. Since the mean aptitude scores of men and women were not significantly different, they have been combined in order to test the difference in the aptitude scores of DPI subgroups.[9] It will be seen in Table 5 that those who had high scores on both projection and reversal (the high projection–high reversal group) had a significantly lower aptitude mean than did the low–low group; and the low projection–high reversal group had lower aptitude scores than the low–low group. This suggests a negative correlation of aptitude and reaction formation. However, since there was not a significant difference in mean aptitude scores of the high projection–low reversal and low–high groups, it may be concluded that aptitude was probably not involved in comparisons of the differential effects of reversal (reaction formation) and projection.[10]

The significant relations (Pearson correlations) of aptitude with other variables were for men: ingroup attitude (−.48), ingroup attitude minus attitude toward self (−.30), the total number of ideas expressed in the Blacky stories (.31), and ingroup deviation scores (−.34). For women, correlation with the total number of ideas was close to significance (.27) and the correlation with opposites associations (−.31) was significant. Partial correlations have been determined for all the significant relationships between these variables (significantly related to aptitude) and other variables, holding aptitude constant. In no case did the correlations in question fall below significance as a result of this partialling procedure.

Interrelations among Opposites Measures: Hypothesis I

The opposites measures to be reported include the following: the frequency of opposites associations, the bimodality index derived from sorting statements about the outgroup, and a number of indices based on Semantic Differential ratings of 24 concepts.

Semantic differential measures. The first step involved analyzing the interscale correlations for each of 24 concepts. That is to say, the ratings of a given concept on each SD scale were correlated with ratings on every other one of the 12 scales. There were consequently 66 correlations for each of the 24 concepts. The results confirmed the close relations among the different evaluative scales[8] already combined as measures of attitude and also demonstrated consistently high correlations between three other scales for each of the 24 concepts. These were the scales designated by the terms: LARGE–SMALL, WIDE–NARROW, and DEEP–SHALLOW, which became the basis of potency ratings. Two kinds of indices were derived from these SD scores: deviation scores and difference scores.

---

[9] Probabilities of .10 or less are reported in all cases. Unless otherwise indicated, these refer to two-tailed probabilities. When variances were not homogeneous, degrees of freedom of t scores were determined by the method described in Dixon and Massey (1951, pp. 104–5).

[10] Aptitude had a correlation of −.16 (men) and −.01 (women) with reversal and −.09 (men) and −.07 (women) with projection. None of these correlations is significant.

a. Deviation scores. The potency and evaluation scores of each person were converted to deviation scores, the distance from the midpoint of the scale to the point of rating without regard to sign. These deviations on the four evaluative scales were summed for each concept to produce the evaluative deviation score for that concept; deviations on the potency scales summed over scales constituted the potency deviation scores. The potency deviation scores and evaluation deviation scores were then correlated over 19 concepts (excluding ingroup and outgroup concepts). Table 1 presents the proportion of 171 intercorrelations of potency scores that were positive and that were significant in a positive direction; the proportion of the 171 intercorrelations of evaluation scores and of the 361 potency and evaluation correlations (19 concepts each with an evaluation and a potency score) that fell into these classes.

It will be seen that 100% of the intercorrelations of potency deviations were positive and 92% and 82% of potency correlations were significantly positive for men and women, respectively. The correlations were somewhat smaller among evaluation scores but they were still highly significant in the positive direction. Even when potency deviations were correlated with evaluation, the proportion of positive correlations was far in excess of chance.

TABLE 1

PERCENTAGE OF POSITIVE AND SIGNIFICANT CORRELATIONS BETWEEN DEVIATION SCORES OF 19 CONCEPTS ON POTENCY AND EVALUATION SCALES

		Evaluation Correlations	Potency Correlations	Evaluation and Potency
Positive	M	98.2%	100%	94.4%
	W	95.9	100	94.4
Significantly Positive*	M	69.5	92.0	31.0
	W	53.8	82.4	42.9
Number r's		171	171	361 Total $N=703$

* $p<.05$.

These results support Hypothesis I for this set of measures by showing that different concepts rated on the same and different scales have a tendency to deviate in the same degree and in the same direction.

In Tables 2 and 3 deviation scores have produced six independent indices: (*a*) average deviations on evaluative scales of concepts that are clearly opposites (Ev_{op}): DIRTY, CLEAN, HONESTY, DECEIT, THRIFTY, WASTEFUL; (*b*) average deviations on evaluative scales of concepts that do not have clear opposites in the list (Ev_{nop}): ME, ENEMY, EARLY, SERVANT, MASTER; (*c*) average deviations on potency scales of concepts in (*a*) referred to as Po_{op}; (*d*) average potency deviations of the concepts in (*b*) referred to as Po_{nop}; (*e*) average deviation on evaluative scales of ingroup concepts; and (*f*) average deviation on evaluative scales of outgroup concepts.

b. Difference scores. It was thought possible that distance between ratings of pairs of opposite concepts might be an appropriate measure of opposites structuring and a general average difference score was calculated, the mean difference in evaluative ratings of six pairs of opposite concepts without regard to sign (Av Diff G).

In addition, four other difference scores have been employed: (*a*) Average difference on evaluative scales between those opposite concepts that might be thought of as having an anal reference (THRIFT–WASTEFULNESS, HONESTY–DECEIT, CLEAN–DIRTY). This score is called Ev_a. (*b*) Difference on evaluative scales between pairs of nonanal opposites (SAFETY–DANGER, SUCCESS–FAILURE, DOMINANCE–SUBMISSION). These are labelled Ev_{na}. (*c*) Difference on potency scales for the concepts in (*a*), called Po_a. (*d*) Difference on potency scales for the concepts named in (*b*) above, called Po_{na}.

c. Interrelations among semantic differential indices. Although the results of

Table 1 suggest a tendency to rate concepts of various kinds similarly on different kinds of scales, it was important to examine possible difference in degree of relationship when different classes of concepts were rated and when concepts were rated on evaluative or potency scales.

When average correlations of difference scores were compared, they were approximately in this order: highest, different concepts rated on the same evaluation scales ($r = .62^*$ and $.53^*$ for men and women, respectively); next highest were correlations of the same concepts rated on potency and on evaluation scales (i.e., on different scales, r's = $.28^*$ and $.32^*$, respectively, for men and women); then, different concepts on the same potency scales (r's = $.31^*$ and .01); finally, different concepts on different scales ($r = .19$ and $-.01$, .16 and .07).

The results were similar for deviation scores. When scales were the same, mean correlations were highest even though the concepts rated were different ($r = .67^*$ and $.68^*$ in men and in women for evaluation scales, and $.74^*$ and $.69^*$ for potency); next were the correlations between indices based on the same concepts rated on different scales ($.40^*$ and $.66^*$ when the concepts were opposites and $.60^*$ and $.65^*$ when they were nonopposites); finally, the lowest correlations appeared when both concepts and scales were different ($.46^*$ and $.40^*$ between potency nonopposites and evaluation opposites and $.33^*$ and $.61^*$ between potency opposites and evaluation nonopposites). In all these correlations the number of cases was equal to 52 for men and 51 for women, and the correlations which were significant are starred. Correlations were converted to z scores before averaging.

Opposites associations. There was evidence of a positive relation between frequency of opposites associations when different stimulus words were used. The number of opposite responses to 100 Kent–Rosanoff words was correlated with the number of opposite responses to 25 other word stimuli. For both men and women this relation was positive and significant ($p < .001$). When the relation of responses to words given on the first day and on the second day (one week later) was determined, the chi squares were again significant and even higher. This makes it appear that a transient set to respond with opposites on some one occasion was not responsible for the correlation.

Bimodality sorts. It should be mentioned at this time that an earlier study (Peak, 1959a) has demonstrated a significant relation in both men and women between the bimodality scores derived from sorting two separate sets of items, one set referring to Negroes and the other set to the members of the ingroup. Correlations between these bimodality scores were .57 and .72 for men and women, respectively, both coefficients being significant at the .01 level.

Interrelations of opposites associations, semantic differential indices, and bimodality indices. Table 2 presents the correlations of each of the indices based on Semantic Differential ratings with the outgroup bimodality index and with the frequency of opposites associations. Since the Semantic Differential deviation scores and difference scores were derived from overlapping data, one *or* the other must be counted in summing up the relations.

In men, 12 of the 13 correlations between Semantic Differential deviation indices and the bimodality index, between the SD deviation indices and opposites associations, and between opposites and bimodality were positive; 5 of the 13 were significant relations. For women 8 of the 13 were positive, none significant. If in this group of indices only those Semantic Differential indices were counted which involved evaluation scales, 9 of the 9 opposite indices were positively correlated in men; 5 of these were significant. In women 6 of the 9 were positive, none significant.

When the Semantic Differential difference indices were counted along with the other opposites indices (bimodality and opposites associations), then 10 out of the 11 correlations were positive for men; 3 of these were significant. In women 8 of 11 were positive, none significant. Once again when potency scores were omitted, 7 out of 7 were positive in men, 3 being significant.

Six out of 7 were positive in women, none significant.

In summary, Hypothesis I was supported clearly by the results from the male sample: a large proportion of the relations among Semantic Differential ratings were significantly positive; relations between bimodality indices were significantly positive, as were relations between frequency of opposite reactions to different lists of words even when tested on different days. Thirty-eight per cent of the relations among bimodality indices, opposites associations, and Semantic Differential indices were significantly positive at the .05 level when SD deviation indices were employed. In women Hypothesis I was supported strongly by the correlations among Semantic Differential measures as well as between the two bimodality sorts and between the two measures of opposites associations. There were no significant positive correlations between the three types of measures (opposites, bimodality, and SD) for women. For both men and women, correlations between measures became smaller when measuring methods differed, when concepts were different, or when different scales were used. Opposites association measures were less clearly related to Semantic Differential measures and

TABLE 2

Generality of Opposites Tendencies–Correlations (Pearson)

Bimodality–Opposites Assoc	M W	.18 (50)[a] .20 (47)[b]			
Bimodality–Outgroup Dev	M W	.32* .12	Opp Assoc–Outgroup Dev	M W	.21 η = .46** −.14
Bimodality–Ingroup Dev	M W	.45** .02	Opp Assoc–Ingroup Dev	M W	.02 .04
Bimodality–Ev_{op} Dev	M W	.36* −.01	Opp Assoc–Ev_{op} Dev	M W	.10 .20
Bimodality–EvNoP Dev	M W	.29* .08	Opp Assoc–Ev_{NoP} Dev	M W	.03 −.03
Bimodality–Po_{oP} Dev	M W	.17 .02	Opp Assoc–Po_{oP} Dev	M W	.15 .08
Bimodality–Po_{NoP} Dev	M W	.24 −.10	Opp Assoc–Po_{NoP} Dev	M W	−.01 −.01
Bimodality–Av Diff G	M W	.31* .03	Opp. Assoc–Av Diff G	M W	.42** .16
Bimodality–Diff Ev_{A}	M W	.37* .01	Opp Assoc–Diff Ev_{A}	M W	.16 .21
Bimodality–Diff Ev_{NA}	M W	.20 −.17	Opp Assoc–Diff Ev_{NA}	M W	.01 .12
Bimodality–Diff Po_{A}	M W	.08 .06	Opp Assoc–Diff Po_{A}	M W	−.16 −.08
Bimodality–Diff Po_{NA}	M W	.01 −.08	Opp Assoc Diff Po_{NA}	M W	.02 .06

[a] N (men) = 52 except for opposites assoc, N = 50.
[b] N (women) = 51 except for opposites assoc, N = 47.
* $p < .05$.
** $p < .01$.

to bimodality measures than the latter two were to each other. Moreover, bimodality indices were less clearly related to potency measures than to evaluation.

Relationships among Defense Preferences: Hypothesis II

Hypothesis II was supported by the fact that there was a significant negative relation between preference for projection and preference for reaction formation or reversal.

Because DPI scores were derived from ranks of five statements made after the presentation of each Blacky picture, each statement representing preference for one of the five defenses, a spurious negative correlation would be expected among the scores. It was necessary, therefore, to find a way of removing this correlation in order to determine the actual relation among DPI preferences.[11] First, the average rank of projection statements and the average rank of those reflecting reversal preference were determined for each person and a scatter plot was made of the relation of the two variables, projection and reversal. Since five ranks were possible this resulted in a 5 by 5 table.[12] This table was then reduced to 4 by 5 categories by making one variable (viz., reversal) assume only four values. Thus, if a person's average projection score was highest (Rank 1) and reversal second, in the reduced table he retained Rank 1 on projection but was also given Rank 1 on reversal; if he had Rank 1 on projection and 4 on reversal, projection remained unchanged but reversal was converted to Rank 3: if projection was Rank 2 and reversal 1, these two ranks were not changed, but reversal Ranks 3, 4, and 5 became 2, 3, and 4, respectively. And so on. In other words, the "reduced" variable took a rank of 1, 2, 3, or 4, columns or rows keeping their original *order* but having ranks independent of those taken by the first variable. A second 4 by 5 table was built from the same data by collapsing the other variable, projection, into four categories in the same fashion, while reversal retained the original five categories.

These four by five tables were analyzed by the Mood maximum likelihood chi square method. For men the chi squares were significant at less than the .05 level for both tables relating projection and reversal; for women the corresponding probabilities were $< .02$ and $< .01$. In order to demonstrate the direction of this relationship the categories were further collapsed to two by two tables. Men and women were combined since the relation between the variables was the same. Phi coefficients for the two tables were $-.44$ and $-.38$.[13] The chi squares from which these coefficients were derived were significant at less than the .001 level.

Avoidance and reaction formation were positively related, the chi square being significant for women but not for men. Preference for projection was positively related to preference for regression ($p < .05$ for men and $< .10$ for women.)

Hostility and Defenses

No specific predictions were made regarding the relation of defense preference and hostility, but it is of interest that projection was positively related to the number of hostility items appearing in the stories (r's $= .34$ and $.32$ in men and women, respectively; significant at the .05 level). Reversal was, on the other hand, negatively correlated with the number of hostility items (r's $= -.37$ and $-.36$, both significant).

It will be explained in the discussion that evidence of several kinds led to an hypothesis that reversal tended to be associated with restriction of psychological distance and difference between certain categories with consequent emphasis on category similarity. It was hypothesized that this might be expected to result in reduced output of Blacky items, because the classification of many ideas as the same would result in fewer total ideas. This proved to be the case, for the total number of ideas produced was negatively related to reversal preference (r's $= -.24$ and $-.31$ with $p <$ .10 for men and $<$ .05 for women.) The relations to projection preference were positive and low ($r = .06$ and .14 for men and for women). The question then arose whether the association of number of hostile ideas with defense preferences was independent of the relation between total ideas produced and defense scores. The positive relation of projection and percentage of hostility items was significant for men ($r = .34$, $p < .05$) and nearly so for women ($r = .26$ with $p < .10$). However, the negative relation of hostility and reversal preference dropped when hostility was expressed as a percentage of the total output (r's $= -.27$ with $p < .10$ and $-.17$ with $p > .10$ for men and women, respectively).

---

[11] John M. Gilbert of the Center for Advanced Study in the Behavioral Sciences suggested this method of handling the problem. We are also indebted to him for help in programming our data for computer runs.

[12] Tied average ranks were assigned to one of the adjacent categories by a table of random numbers.

[13] With the given marginals, maximum phi for the first table was .940; for the second, .867.

Table 5 reveals these same differences somewhat more clearly. This Table represents the mean scores of the variables within groups resulting from dividing the sample simultaneously on the basis of projection and reversal preferences derived from hostility pictures. Men with low projection and high reversal preferences produced significantly fewer total ideas than did those with low scores on both ($p <$ .05); also, those with low projection and low reversal scores produced significantly more ideas than those with high scores on both ($p < .01$). High–low scorers were more productive than were those high on both projection and reversal, but not significantly so. The *percentage* of hostility items was significantly greater in high projection–low reversal conditions than in the low–high condition ($p < .05$) and the high–low group also produced a greater percentage of hostility items than did the low–low group ($p < .10$). This analysis, like that based on the correlation coefficient, suggests a closer relation between per cent hostility and projection than between per cent hostility and reversal.

Women high–high scorers produced fewer ideas than did those with high–low scores ($p < .05$); high–low scorers expressed more than those low on projection and high on reversal ($p < .01$). Results for the percentage of hostility items were in the same direction as for the male sample but the differences between group means were not significant for women.

Defenses and Opposites: Hypothesis III

The relations between opposites measures and defense preference have been analyzed by several different methods. There were, first of all, the different opposites indices to be taken into account, as well as different indices of defense preference. In the second place, the relations between defense preference and opposites scores have been expressed in two ways: in terms of Pearson correlation coefficients and by comparing the means of opposites indices in the four groups resulting from dividing the sample simultaneously on reaction formation scores and on projection scores. Only the comparison of means is reported at this time. Correlation coefficients were not markedly different in their implications from differences in means of the groups, but where there were differences the means are probably more accurate because the level of reversal and projection scores was controlled for each comparison.

Tables 3 and 4 present the means for the four DPI groups resulting from dividing the sample simultaneously on reversal and on projection. In Table 3, DPI indices from hostility pictures were used; in Table 4, the indices were those derived from reactions to those Blacky pictures which showed evidence of conflict for the individual when the story written about the picture was scored for conflict by the Blum (1950) method.

It was predicted in Hypothesis III that high projection scores would be associated with high opposite scores and high reversal scores with low opposite scores. Therefore, those high on projection and high on reaction formation should have higher opposites scores than the low–high DPI group, projection varying and reversal constant and high. Moreover, high–low scorers should have larger opposites scores than low–low scorers, indicating the effects of projection with reaction formation constant and low. On the other hand, those with high projection and high reversal preference should have lower opposites scores than the high–low group, and the low–high group should have smaller scores than the low–low *S*s, both of these comparisons indicating the effects of reversal preference with projec-

TABLE 3

OPPOSITE SCORES OF DPI GROUPS
(Hostility Pictures)

		High Proj–High R I		High Proj–Low R II		Low Proj–High R III		Low Proj–Low R IV		Significance and Direction of Difference
		Mean	*s*	Mean	*s*	Mean	*s*	Mean	*s*	
Bimodality	M F	15.8 20.2	5.2 7.9	17.2 16.5	5.6 6.4	14.9 15.3	7.0 4.4	14.5 13.2	3.1 3.0	I>IV**
Opp Assoc	M F	27.3 20.8	13.9 11.6	21.1 22.6	8.8 13.6	26.3 23.4	8.7 12.3	27.9 26.5	9.7 10.5	II<IV*
OG Dev	M F	21.4 33.7	7.9 14.4	30.7 35.9	12.3 12.3	27.5 29.8	16.1 10.5	25.9 29.0	14.5 7.6	I<II*
IG Dev	M F	36.9 52.8	12.9 6.3	47.1 51.3	9.4 7.7	40.9 47.6	11.4 9.1	39.7 49.7	13.9 9.1	I<II**
Evopp Dev	M F	41.0 52.2	11.3 6.1	49.0 51.3	6.3 7.4	43.4 46.7	9.1 9.6	44.2 48.2	8.9 7.1	II>III**
EvNoP Dev	M F	24.9 34.3	7.5 8.3	32.4 34.8	5.8 7.6	28.7 32.2	8.4 8.5	30.3 32.2	8.4 5.5	I<II**
Po$_{op}$ Dev	M F	25.0 30.8	6.7 8.6	26.1 31.3	8.7 9.5	22.1 29.3	9.8 8.3	29.9 33.2	6.5 7.5	III>IV***
Po$_{NoP}$ Dev	M F	18.3 22.2	6.1 8.3	19.9 22.2	6.4 7.2	17.1 21.3	6.5 5.8	20.0 24.2	3.6 6.3	
Ev$_{NA}$	M F	18.7 24.3	7.2 4.0	23.7 24.6	7.2 7.4	20.2 24.1	5.3 7.1	20.9 23.7	4.2 6.4	
Ev$_{A}$	M F	26.3 34.5	8.4 4.2	29.9 33.2	6.5 6.3	28.3 29.7	6.8 7.4	28.9 30.8	6.0 6.7	
Po$_{NA}$	M F	4.7 4.5	2.4 2.2	5.8 6.5	3.0 3.0	6.0 7.1	1.5 2.7	5.8 8.2	2.7 1.9	I<IV***; I<III**
Po$_{A}$	M F	2.6 4.2	1.1 1.0	3.6 2.9	2.2 1.6	3.2 3.8	1.6 1.8	3.5 2.8	1.7 1.8	I>II**
N	M F	7 6		16 17		17 20		10 6		

* $p < .10$.
** $p < .05$.
*** $p < .01$.

tion held constant. Finally, the group with high projection scores and low reversal scores should have higher opposite scores than the one with low–high scores, both defense preferences varying in opposite directions.

The relative size of the opposites scores of the above pairs has been determined for men and for women, for each opposites measure. The number of differences in the predicted direction has been counted for two sets of comparisons, the first using Semantic Differential indices derived from *deviation* scores, bimodality scores, and opposites associations, and the second including the Semantic Differential *difference* scores along with bimodality and opposites associations. These counts indicate the proportion of the comparisons of mean opposites scores for the DPI groups with different defense preferences which were in the direction predicted by Hypothesis III. In the section following, DPI indices based on hostility pictures were used to determine groups; in the next section after that, indices derived from conflict pictures were employed.

Groups determined by hostility DPI's. In the first count using *deviation* scores, bimodality, and opposites associations *women* showed no significant predicted differences in mean opposite scores of the different DPI groups. However, 12 of 16 comparisons in which projection was varying, with reversal constant, were in the predicted direction. By the sign test this number of differences in one direction would occur by chance less than five per cent of the time ($p = .038$, one-tailed). With reversal varying, there was not a significant number of differences in the predicted direction. For the comparison of the high projection–low reversal group with the low projection–high reversal group seven of eight differences in means were as predicted ($p = .035$, one-tailed).

For the *male* sample using *deviation* scores, bimodality, and opposites associations, 5 of the 16 comparisons indicated the differences in opposites scores to be significant in the predicted direction, and in all of these significant comparisons, reversal was varying. Twelve of the 16 comparisons between high and low reaction formation means with projection constant were in the predicted direction ($p = .038$, one-tailed). Nine of 16 were in the predicted direction with projection varying and reversal constant. Comparing the high projection–low reversal means with low–high means, seven of eight were as predicted ($p = .035$, one-tailed).

On the second count, including *difference scores* for measures involving Semantic Differential ratings, plus bimodality and opposites associations, the results were in the same direction but less clear. Only when the average difference score was the measure was there a significant predicted difference in a Semantic Differential measure in men. One such measure was significant for women, Po_{na}. In men, 6 of 12 comparisons were in the predicted direction with projection varying; 9 of 12 with reversal varying ($p = .073$, one-tailed). In women, none of the frequencies exceeded chance expectation, when difference scores were involved.

Groups determined by conflict DPI's. When the sample was divided on the basis of DPI's calculated only from the pictures on which an *S* showed evidences of conflict, then the results changed in two ways (Table 4). The relations of DPI scores to the opposites measures became somewhat more significant and the results for men and women were more alike in that there was no difference in the magnitude of the effect of projection and reversal on opposite scores for men and women. The results for men and women have been combined, therefore, in testing the frequencies of relations in predicted directions.

Using the *deviation* measures, bimodality scores and opposites association scores, 23 of 32 comparisons were as predicted when projection was varying and reversal constant ($p < .02$); 24 of 32 were as predicted when reversal was varying and projection held constant ($p < .01$). When men and women were treated separately, $p = .038$ and .105 with projection varying; with reversal varying, $p = .011$ and .105 for men and women, respectively. In 7 of 8 comparisons ($p = .035$) for men and 6 of 8

($p = .145$) for women, high projection–low reversal means were greater than low projection–high reversal means.

With *difference* scores replacing deviation scores the result were in the same direction, but the frequency of the differences in the predicted direction did not exceed chance.

In women there were four significant differences in the predicted direction out of 16 comparisons: bimodality scores were larger for the high–projection low reversal group than for the high–high group, scores for the high–low group were greater than for low–lows and scores for the high–low group were greater than for the low–high

TABLE 4

OPPOSITE SCORES OF DPI GROUPS
(Conflict Pictures)

		High Proj– High R I		High Proj– Low R II		Low Proj– High R III		Low Proj– Low R IV		Significance and Direction of Differences
		Mean	*s*	Mean	*s*	Mean	*s*	Mean	*s*	
Bimodality	M	13.9	3.5	16.2	6.0	15.5	6.5	15.7	5.2	I<II*; II>III*; II>IV**
	F	15.2	4.6	18.9	6.0	14.9	6.0	14.0	4.2	
Opp Assoc	M	26.7	8.9	24.2	12.5	27.2	10.9	20.4	10.9	I>II*; I>III***
	F	29.5	6.4	23.6	11.6	20.8	12.9	22.2	15.4	
OG Dev	M	23.0	11.5	30.1	11.8	24.1	14.9	28.6	16.9	
	F	29.3	11.0	35.6	9.6	31.9	14.9	31.4	9.3	
IG Dev	M	41.5	13.2	44.4	9.9	36.4	11.4	49.6	9.9	II>III**; III<IV*** I<IV*
	F	47.1	9.5	50.3	7.5	49.5	9.5	53.2	4.7	
$\mathrm{Ev_{opp}}$ Dev	M	46.5	7.8	48.2	7.8	38.9	8.1	47.9	8.9	I>III**; II>III***; III<IV**
	F	49.2	7.1	50.0	7.8	48.5	10.1	49.0	8.9	
$\mathrm{Ev_{NoP}}$ Dev	M	28.1	4.3	32.8	6.4	25.3	8.6	31.1	7.5	I<II**; II>III***; III<IV*
	F	33.2	8.5	34.1	7.5	33.0	8.7	33.2	6.3	
$\mathrm{Po_{OP}}$ Dev	M	28.1	8.4	28.9	8.1	21.9	8.1	21.5	7.7	I>III*; III<II***; II>IV**
	F	31.6	8.5	29.8	9.3	30.9	8.7	31.0	8.3	
$\mathrm{Po_{NoP}}$ Dev	M	18.7	5.2	20.9	5.4	16.6	6.3	17.2	5.1	II>III**
	F	23.0	6.5	20.4	7.5	22.7	6.5	23.1	4.5	
$\mathrm{Ev_{NA}}$	M	21.6	4.9	23.1	6.7	17.2	5.9	23.2	3.1	I>III*; II>III***; III<IV** I<IV*
	F	22.2	5.7	23.9	6.7	24.9	7.0	27.4	7.2	
$\mathrm{Ev_{A}}$	M	30.5	5.4	30.1	7.2	25.2	6.1	30.0	6.8	I>III**; II>III**
	F	31.5	5.9	32.6	6.2	31.1	8.8	31.1	5.5	
$\mathrm{Po_{NA}}$	M	7.4	3.2	5.3	2.7	5.1	1.5	6.2	2.8	I>III* I<III*; I<IV*
	F	5.5	2.6	6.6	2.9	7.5	3.2	7.5	1.9	
PoA	M	3.4	2.1	3.2	1.8	3.3	1.7	3.5	1.9	
	F	3.8	2.6	2.9	1.5	3.3	1.8	4.0	1.2	

* $p < .10$.
** $p < .05$.
*** $p < .01$.

group; the number of opposite associations in high-high scores was greater than in the low–high group. In men, 12 of 16 comparisons were significant in the predicted direction when deviation indices were involved and 6 of 12 with difference scores.

It must be concluded that Hypothesis III, though supported in part, needs certain qualifications. It is necessary to sample more systematically the content areas tapped by the various opposites measures described in this section in order to determine what exactly was responsible for the higher association of the opposites series with defense preferences expressed in areas of conflict than in response to hostility pictures. It appears likely that preference for reversal is associated with low opposites scores and projection with high ones when preferences are determined in situations comparable to those involved in the opposites measures. This is the plausible explanation of the closer relation between ingroup–outgroup difference and DPI scores based on hostility pictures; both the attitude measures and hostility pictures presumably produced reactions to hostility. It can only be surmised that the general opposites measures of Tables 3 and 4 activated categories that overlapped with conflict pictures to a greater degree than with hostility pictures.

A second qualification is one that must be stated for almost every hypothesis in this study. Hypothesis III was more strongly supported for men than for women, though tendencies were in the same direction for both.

Finally, a comment should be made on the more significant relations obtained with SD deviation scores than with SD difference scores. Since the latter were derived from differences in Semantic Differential ratings of pairs of concepts assumed to be opposites, a tendency to extreme reactions to both concepts would be obscured if *S*s did not place the two concepts on opposite sides of the scales. This tended to be the case, for example, for the pair SAFETY-DANGER. Men, in particular, did not always rate DANGER as bad, in which case the difference between the DANGER and SAFETY score tended to be small, even when both scores were extreme. This suggests the need for further attention to the nature of the structures being analyzed.

Defense Preference and Attitude: Hypotheses IV and V

Here, as in other sections, relationships have been analyzed by correlating the variables and by comparing the means of portions of the sample which have known levels of preference for reversal and projection. The latter analysis is reported at this time.

Table 5 presents means of the important variables for each of the following groups of *S*s: high projection–high reversal, high projection–low reversal, low projection–high reversal, and low on both defenses. Figure 1 indicates the means of the attitude variables in men and women for each of these DPI groups. The horizontal lines represent evaluative scales with the mean ingroup (family, friends, Americans), ME, and outgroup positions (Negro neighbors, Negro roommate, foreigner) for each DPI group. The attitude positions of different groups are joined by solid lines for men and by dotted lines for women.

The significant differences in these variables will be indicated in the succeeding paragraphs and their implications discussed in the final section.

Ingroup attitudes. Only one comparison of ingroup attitudes in the different DPI groups reached an acceptable level of sig-

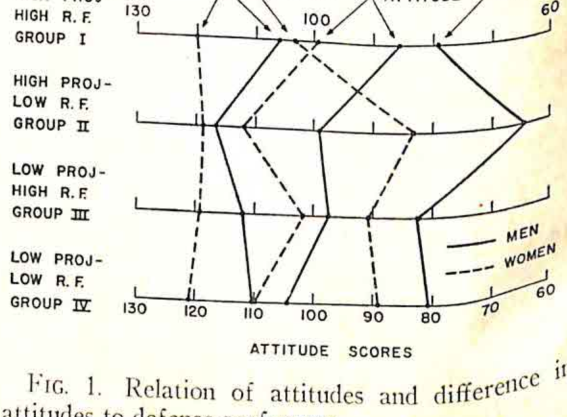


FIG. 1. Relation of attitudes and difference in attitudes to defense preference.

nificance: in women with reversal constant and high, there were more favorable attitudes toward the ingroup when projection was high than when it was low ($p < .05$), suggesting the predicted effect of projection on ingroup attitude. In men with projection constant and high, there were more favorable attitudes toward the ingroup when reversal preference was low than when it was high ($p < .10$), which implies that reversal was making the ingroup attitude less favorable.

TABLE 5

Averages of Variables as Function of Reversal and Projection
(Hostility Pictures)

		High Proj–High R I		High Proj–Low R II		Low Proj–High R III		Low Proj–Low R IV		*P*'s between Variables Indicated
		Mean	*s*	Mean	*s*	Mean	*s*	Mean	*s*	
Ingroup Attitude	M	106.6	14.7	117.2	10.6	112.2	12.0	110.5	15.2	I<II^{a}
	W	125.8	6.2	123.2	7.8	119.0	9.0	120.3	11.7	I>III^{b}
ME Attitude	M	85.7	22.6	99.3	19.3	97.4	18.1	104.4	16.5	
	W	99.9	15.1	111.9	14.2	101.7	17.3	105.0	13.3	I<II^{a}
Outgroup Attitude	M	79.4	13.1	63.3	23.2	82.7	25.1	80.2	18.5	I>II^{b}; II<IV^{d}; II<III^{b}
	W	103.0	16.2	82.8	27.8	90.5	22.7	89.5	11.8	I>II^{b}
|IG–ME| Attitude	M	24.6	13.9	20.2	16.1	19.4	11.1	10.7	8.0	I>IV^{b}; III>IV^{d}
	W	25.8	14.2	14.4	10.5	20.7	14.0	15.3	8.4	
	C	25.1	13.5	17.1	13.4	20.1	12.1	12.4	8.2	I>II^{a}; I>IV^{d}; III>IV^{d}
|ME–OG| Attitude	M	24.0	14.2	36.2	22.6	25.5	18.1	26.2	25.2	
	W	14.3	10.1	31.0	25.6	24.1	22.0	23.5	19.5	I<II^{b}
	C	19.5	13.0	33.4	24.1	24.8	20.1	25.2	22.8	I<II^{c}; II>III^{a}
|IG–OG| Attitude	M	27.1	18.7	54.3	23.8	30.2	21.6	30.3	22.4	I<II^{d}; II>III^{d}; II>IV^{c}
	W	26.1	15.4	38.7	27.4	33.3	21.9	30.8	21.4	
	C	26.7	16.6	45.9	26.7	31.9	21.5	30.5	21.5	I<II^{d}; II>III^{c}; II>IV^{b}
Aptitude	M	120.7	14.0	125.7	18.5	123.1	21.7	132.6	18.7	
	W	113.6	15.4	123.9	20.2	119.8	20.4	134.0	17.7	
	C	117.4	14.4	124.7	19.3	121.3	20.7	133.1	17.8	I<IV^{b}; III<IV^{b}
% Hostility Ideas/Story	M	4.6	2.0	4.8	1.9	3.6	1.1	3.8	1.0	II>IV^{a}; II>III^{b}
	W	3.8	2.1	3.6	1.3	3.1	1.3	2.7	0.5	
No. Ideas/Story	M	20.2	2.7	23.8	7.0	21.9	4.2	24.4	2.0	I<IV^{d}; III<IV^{b}
	W	23.2	5.2	28.1	3.7	24.4	4.1	28.0	3.9	I<II^{b}; II>III^{d}; III<IV^{a}
F Score	M	59.6	15.9	70.6	9.0	67.3	10.0	64.1	16.7	I<II^{a}
	W	52.5	9.8	67.9	10.3	60.5	10.5	66.2	13.1	I<II^{d}; I<III^{a}; II>III^{b}; I<IV^{a}
No. Cases	M	7		16		17		10		
	W	6		19		20		6		
	C	13		35		37		16		

[a] $p < .10$.
[b] $p < .05$.
[c] $p < .02$.
[d] $p < .01$.

Outgroup attitudes. In both men and women with projection constant and high, outgroup attitudes were significantly more favorable in those with high reversal preference than in those with low ($p < .05$), which was predicted. In men with reversal constant and low, those with high preference for projection had a significantly less favorable attitude toward the outgroup than did those with low projection scores ($p < .01$). The difference was in the same direction for women but did not reach an acceptable level of significance. The high projection–low reversal group had significantly less favorable outgroup attitudes than did the low projection–high reversal group of men ($p < .05$). Again, a difference in the same direction for women was not significant. These differences were in the predicted direction according to Hypotheses IV and V.

There were no significant differences in attitude toward the self-concept (ME) as a function of preference for reversal or projection, though in both men and women high–high scorers were less favorable to ME than high–low scorers ($p < .10$ for women); that is, with projection constant and high there tended to be more favorable self-attitudes associated with low reversal than with high reversal preference.

Ingroup–outgroup attitude differences. When outgroup attitude ratings were subtracted from ingroup ratings for each *S* and the differences averaged within each DPI group without regard to sign, both men and women showed maximum difference in the group with high preference for projection and low preference for reversal and less difference for the high reversal groups.

Scores of men and women have not been combined in some analyses because of significant differences in means of the two samples. However, differences between men and women were not significant for the mean ingroup–outgroup attitude differences, for mean ingroup–ME attitude differences, or for mean outgroup–ME attitude differences, as indicated in Table 6. Therefore, the two samples have been combined in analyzing these difference scores (C in Table 5).

The association of small ingroup–outgroup differences with reversal preference was demonstrated in the significant difference between the high reversal–high projection group and the low reversal–high projection group ($p < .01$). However, there was not a significant difference in the high reversal–low projection and low–reversal low projection groups. In short, the effects of reversal preference were clear only with projection preference high and constant. The comparison of the high projection–low reversal group and the low projection–high reversal group showed a significant difference in the expected direction ($p < .02$), a difference which cannot be attributed exclusively to reversal or to projection. The significant difference in the high projection–low reversal and low projection–low reversal groups ($p < .05$) suggests that with reversal constant and low, high preference for projection favors larger ingroup–outgroup differences. On the other hand, the expected effects of projection preference did not appear when reversal was constant and high and projection varied. These results suggest that when both reaction formation and projection were favored defenses, the former obscured the effects of the latter, so that effects of projection appeared only when reversal preference was low.

ME–outgroup differences. The mean differences in this pair of attitudes followed the same pattern as the ingroup–outgroup differences, and for the combined sample the same comparisons of DPI groups were significant except that the comparison of the high projection–low reversal group with low projection–low reversal group did not reach a significant level.

Ingroup–ME differences. Since there were significant positive correlations between attitudes toward the concept ME and toward the ingroup (Pearson r's = .34 and .45 for men and women, respectively), it was not surprising to find that these attitudes tended to change together from one DPI group to another. There were, however, two significant differences for the total sample. The low projection–high reversal group showed greater ME–ingroup distance than the low–low group ($p < .01$). This comparison reflects the association of reversal preference with increased difference in self- and ingroup ratings when compared to low reversal. No significant effects of projection on ME–ingroup differences were revealed in the group comparisons though high projection groups also showed somewhat greater ME–ingroup differences than did low projection groups when reversal was constant.

The combination of high reversal and high projection preference produced larger ingroup–ME differences than did any other condition and low preference for both defenses produced the smallest differences ($p < .01$ between these two groups). In other words, self-attitudes and ingroup attitudes were farther apart with combined preference for projection and reversal than with preference for neither.

Analysis of Separate Ingroup and Outgroup Concepts

Ingroup attitudes were defined for our purposes as a combination of attitudes toward family, friends, and Americans. Outgroup attitudes included reactions to Negro neighbor, Negro roommate, and foreigners. Predictions about the composite indices have been tested and reported. It is important, however, to examine two further questions: (*a*) Are the results characteristic of each component of the ingroup and outgroup indices? (*b*) Are there differences in the organization of ingroup components and of outgroup components as a function of defense preference? No predictions were made about this second point but it is clear that intracategory relationships must be described in any complete account of structures. For example, Bruner, Goodnow, and Austin (1956) and Pettigrew (1958) have reported individual differences in category breadth and others have assumed that prejudiced persons ignore distinctions within the outgroup. The present data provide information only about evaluative ratings of components of ingroup and outgroup categories, but it is of interest to note relations among these properties even though they are incomplete aspects of total ingroup–outgroup structure.

The relation of defense preference to differences in component ingroup–outgroup concepts. Mean attitude ratings for each of the six concepts listed above as well as mean difference in ratings of the concepts without regard to sign have been determined within each DPI group.

Relative distances between the attitudes involving Negroes (roommate and neighbors) and each of the ingroup concepts within DPI groups were those reported for the composite indices. In every one of the 12 comparisons (6 for men and 6 for women) the greatest Negro–ingroup distance was shown in the high projection–low reversal group.

In men the same was true of the three comparisons of ratings of FOREIGNER and each of the ingroup concepts. However, this was not the case for the women *S*s who did not react to the concept FOREIGNER as opposite to the ingroup concepts. In the first place, the mean rating of this concept was related to DPI scores as were ingroup concepts, tending to become more favorable in the high projection groups and less favorable with high reversal preference. And, secondly, women with high reversal preference showed the greatest difference between FOREIGNER and AMERICAN, FOREIGNER and FAMILY, and FOREIGNER and FRIEND. This was the typical pattern of relations between ingroup concepts, as will be indicated in the next section.

In short, the a priori assignment of the concepts in question to ingroup and outgroup categories appears to have been justified in the case of men. For women attitudes toward the concept of FOREIGNER were evidently not operating like the attitudes toward the concept of the NEGRO.

Within-category distances. Table 6 indicates that differences were smallest between FRIEND and FAMILY, NEGRO ROOMMATE and NEGRO NEIGHBOR, AMERICAN and FAMILY, and AMERICAN and FRIEND, all within-category pairs. The FOREIGNER–NEGRO ROOMMATE and FOREIGNER–NEGRO NEIGHBOR pairs were also thought of as involving within-category comparisons, but the differences between those pairs of attitudes were slightly larger than the distance be-

TABLE 6

MEAN DIFFERENCES IN ATTITUDES
Total Sample ($N = 101$)

Pair	Mean	s
|Friend-Family|	1.9	2.5
|Neg Rm-Neg N|	5.1	4.8
|Amer-Family|	5.3	6.3
|Amer-Friend|	5.7	5.1
|For-Amer|	6.2	5.2
|For-Family|	8.1	6.2
|For-Friend|	8.3	2.0
|Neg N-For|	9.7	7.7
|Neg R-For|	9.7	8.5
|Neg N-Amer|	12.6	10.0
|Neg R-Amer|	12.7	10.6
|Neg N-Family|	15.7	10.0
|Neg R-Family|	15.8	11.0
|Neg R-Friend|	16.0	11.0
|Neg N-Friend|	16.4	10.0

tween FOREIGNER and AMERICAN, FOREIGNER and FRIEND, and FOREIGNER and FAMILY. Here again the concept of FOREIGNER was clearly marginal as an outgroup concept. Nevertheless, distances were in general smaller between components of the ingroup as originally defined and between components of the outgroup than across these categories, except where the FOREIGNER concept was involved.

When distances between ingroup attitudes were examined as a function of defense preferences, the following results were obtained. There were no significant differences between DPI groups in mean FAMILY and FRIEND attitude differences. However, the mean attitude difference between AMERICAN and FAMILY was significantly larger ($p < .01$) in the low projection–high reversal group than in the high projection–low reversal group. Differences between AMERICAN and FRIEND attitudes were also significantly larger ($p < .02$) for the low projection–high reversal group. The mean of the three ingroup differences was significantly larger for those preferring reversal than for projectors ($p < .05$).

The mean difference in attitudes toward NEGRO ROOMMATE and NEGRO NEIGHBOR was not significantly related to defense preference. However, the NEGRO ROOMMATE–FOREIGNER comparison and the NEGRO NEIGHBOR–FOREIGNER comparison showed relations to DPI that were more like those we came to expect in comparisons *between* categories; i.e., attitude differences in the first pair were greater for the high projection–low reversal *S*s than for low–high *S*s ($p < .10$), and the high projection–high reversal group had greater mean difference in these attitudes than did the high–low group ($p < .05$). Mean attitude differences between NEGRO NEIGHBOR and FOREIGNER were also significantly different ($p < .05$) between high–lows and low–highs, the first group showing the larger mean difference.

In summary, attitudes toward the three ingroup concepts behaved in a similar manner and there was a tendency for high projectors to have smaller distances between ingroup concepts, while those favoring reversal had greater ingroup differences. For the three outgroup concepts, relations between NEGRO and FOREIGNER had the characteristics of between-category relations such as have been reported for ingroup–outgroup composite indices. High projection preference was related to greater differences and reversal to smaller differences. Since NEGRO ROOMMATE–NEGRO NEIGHBOR differences showed no relation to DPI scores, no generalization can be made about mean outgroup attitude differences as a function of DPI scores.

F Scores, Opposites, and Defense Preference

Scores on the California F Scale were not significantly related to any of the general opposites measures. They were, however, significantly related to a difference in ingroup and outgroup attitudes in both men and women ($p < .05$ for men and $< .01$ for women).

High F scores were associated with high preference for projection and with low preference for reversal. In Table 5 it will be seen that the mean F score of the high projection–low reversal group was higher than that of any of the other groups and was significantly higher in women than the low projection–high reversal group ($p < .05$) or than the high projection–high reversal group ($p < .01$) which had the lowest F score in both men and women. In men the high–low group had higher F scores than did the high–high group ($p < .10$). This negative relation of F to reversal would not have been expected from theories of authoritarian personality structure, which have assumed a tendency to reversal preference as characteristic of the high F individual (Adorno et al, 1950).

Moreover, the fact that the combination of high projection with high reversal was associated with the *lowest* F scores in both men and women suggests as already noted that reversal may be a defense against the tendency toward projection in this group. The effects of reversal seem to have been dominant here as they were in determining ingroup–outgroup distance.

In other respects the F score results were the familiar ones. High F's had more favorable ingroup and less favorable outgroup attitudes as well as a greater distance between these attitudes. The higher the F score, the smaller was the ingroup–ME distance and the greater the ME–outgroup distance.

Sex Differences: Hypothesis VI

The primary reason for analyzing the difference in reactions of men and women was to determine whether the samples could be combined. There was a sufficient number of significant differences to suggest that separate analyses should be made. As already indicated, in those cases where the differences were not significant, the samples were pooled.

Table 7 compares the average scores of men and women on the principal variables in the study. The following differences were significant at the .05 level or less. Ingroup, outgroup, and self-attitudes were more positive in women than in men. Of 13 opposites indices all but one showed women to have larger opposites scores than men and for that exception (opposites associa-

TABLE 7

MEAN SCORES MEN AND WOMEN

	Men			Women			*P* Difference
	Mean	*s*	*N*	Mean	*s*	*N*	
Outgroup Attitude[a]	75.8	21.9	52	89.0	23.9	51	<.01
Ingroup Attitude[a]	112.8	12.5	52	121.5	8.7	51	<.001
ME Attitude[a]	98.4	18.9	52	105.6	15.6	51	<.05
|Ingroup-Outgroup|	37.5	24.5	52	34.1	23.4	51	>.10
|ME-Outgroup|	28.9	20.8	52	25.5	22.1	51	>.10
|Ingroup-ME|	18.7	13.2	52	18.3	12.2	51	>.10
Bimodality	1.2	5.6	52	6.1	5.8	51	<.001
|Av Diff|	25.9	6.7	52	29.1	5.7	51	<.01
Opposites Assoc	32.6	13.1	50	32.6	14.7	47	>.10
Deviation Outgroup	27.0	13.6	52	32.4	11.5	51	<.05
Deviation Ingroup	42.3	11.5	52	49.9	8.3	51	<.001
% Hostile Ideas	4.3	1.6	51	3.4	1.4	51	<.01
No. Ideas/Story	22.7	4.9	51	26.2	4.7	51	<.001
F score	65.9	12.8	52	63.0	11.5	51	>.10
Reversal[b]	16.5	3.3	51	14.5	3.8	51	<.01
Projection[b]	14.5	3.9	50	16.0	4.1	51	<.10
Aptitude	125.0	18.9	48	122.5	19.8	49	>.10

[a] High score, favorable attitude. Divide by 12 to determine average scale position on 11 point scale.

[b] High score, low preference. Divide by 5 to determine average rank. These scores are derived from the Hostility pictures.

tions) men and women had equal scores.[14] On 10 of the 13 opposites measures the women's scores were significantly larger than the men's at the .05 level.

A similar tendency to extreme reactions in women has been reported by Osgood, Suci, and Tannenbaum (1957, p. 234) and by Shapiro and Tagiuri (1959). It is to be noted, however, that the average distances between ingroup and outgroup, ME and outgroup, ME and ingroup were not significantly different in men and women.

Men showed a higher percentage of hostility items ($p < .01$); women produced a greater number of ideas on the Blacky stories ($p < .001$). Men gave higher preference ratings to projection than did women ($p < .05$, one-tailed predicted) and women were higher on reaction formation ($p < .01$). This last item supports Hypothesis VI.

Perhaps the most striking sex difference in this study is the fact that relations between variables were in general much stronger in men than in women, though in practically all cases results were in the same direction. Only relations of F scores to the other variables were clearer in women. But again, the results were uniformly in the same direction for men and women.

It is not possible to be sure of the reason for this result but two hypotheses suggest themselves. First, it may be that the DPI measures were more valid for men than for women. It has been suggested, for example, that *S*s tend to interpret the name *Blacky*, the central figure of the DPI test of that name, as a male and that this makes it more difficult for women to identify with Blacky. On the other hand, the F scale was first standardized on women and may possibly be somewhat better adapted to them.

In the second place, it may be that women have less clearly crystallized and organized concepts in the areas which we have studied; this would result in more independence of reactions to different situations. Such a theory is compatible with the observation that there were consistent relations between opposites measures in women only *within* rather circumscribed areas; viz., between Semantic Differential reactions, between different bimodality scores, between opposites scores based on different lists responded to on different occasions, but very little consistency across these different opposites situations.

DISCUSSION

Characteristics of the defenses of reversal and projection, which have been observed in this study, must now be summarized and interpreted. High preference for *reversal* as a defense was associated with small differences between ingroup and outgroup attitudes, between ME and outgroup attitudes, and under certain conditions with low opposites scores. There was also a positive relation between this defense preference and favorable outgroup attitudes, less favorable self-attitudes, low projection scores, low F scores, low frequency of hostility items, and of total output of items in the stories written about the Blacky pictures. The difference in self- and ingroup attitudes was significantly greater when reversal preference was high and projection constant and low. Attitude distances between the three components of the ingroup concepts (FAMILY, FRIENDS, AMERICAN) were larger in the low projection–high reversal group than in the high projection–low reversal group.

On the other hand, a preference for *projection* tended to be related to greater differences between ingroup and outgroup attitudes, to more favorable ingroup attitudes, to less favorable outgroup attitudes, to a high percentage of hostility responses, to high F scores, and to high opposites scores under certain conditions. Projectors tended to rate ingroup concepts closer together on evaluative scales.

Styles of Dealing with Psychological Distance

An effort has been made to incorporate these results into a general theory in which psychological distance has been employed as a central construct in dealing with the problems of motivation (Peak, 1955), perception (Peak, 1958d), attitude change, and generalization (Peak, 1958a, 1958b, 1958c).

---

[14] Only five opposites measures have been entered in the Table, but all were compared.

In the motivation article (Peak, 1955) a motive state has been described as an activated structure consisting of two or more related events which are (*a*) close enough together in psychological space, by virtue of sharing a common dimension or because of associative linkage, to result in a relatively high probability of mutual activation of one event by arousal of the other, but (*b*) sufficiently disparate so that the two events are not reacted to as the same. For example, a system including the painful thought of the "D" I made on the last quiz and the pleasant thought of the grade of "A" may have such a relation. As long as the appropriate amount of disparity and association exists between the concept of "D" with its negative affect and the concept of "A" with its more positive affect, there will be a tendency for the arousal of one of these events to result in the arousal of the other one, with resulting activation of overt behavior. Moreover, this activity will tend to persist until the distance between the events is either increased or decreased markedly. Thus, anything which reduces the disparity between the terms of such a structure to the point where they are reacted to as the *same* will remove the motivational activity. This could happen, for example, if I actually received an A on an examination. It would also happen as a result of reacting defensively to the A and the D states as though they were not different. On the other hand, whatever increases the distance from the A state to the perceived state, like making an E on an examination, may increase the distance between the terms of the system to the point where one event, the thought of the grade of "E," no longer elicits the other, the thought of "A" with sufficiently high probability to maintain activation in the system. Such an increase in disparity may also be accomplished defensively and unrealistically by changing the perceived position of one or both terms of the disparity in such a way that their difference is exaggerated.

This, then, implies two basically different ways of reacting which may constitute defensive adjustments. In the one case, the distance or difference between two terms of a motive structure is reduced, causing activity which stems from this motive system to cease. The results of the present study suggest that those preferring reversal may have developed a style of reacting in this way to disparity between ingroup and outgroup attitudes. In the other case, psychological distance is presumed to increase in a manner compatible with the behavior of projectors in this study. It would be assumed that these styles of adjusting have been learned as methods of reducing disparity in particular situations and that they have become the favored styles of reacting where disparity could not be reduced directly by reaching goals.

When the results were viewed in this framework it became apparent that the theory had other implications which could be checked in our data. We have found reversal to be negatively related to the number of hostile items in the Blacky stories and projection related positively to those items. It was not at once apparent how these findings could be explained in terms of styles of dealing with distance. Then, the following hypothesis occurred to us. The number of hostility responses would be increased by greater total production of ideas; conversely, these scores would be decreased by restrictions on output. It also appeared reasonable to assume that a student who is writing stories for two minutes will usually operate under a self-administered set which prevents him from repeating the *same idea* within this brief period. Presumably, the person who tends to neglect differences will regard more ideas as the same. Consequently, such a person (with reversal preferences) would have fewer *different* ideas available and would consequently tend to produce less material within the two-minute period. This would result in lower hostility scores.

As already noted, the emergence of this hypothesis, after the data were collected and the main analysis completed, led us to count the number of "ideas" produced in the course of writing the projective stories and the results tended to support the hypothesis by revealing a smaller number of ideas produced by those preferring reversal ($p < .10$ for men and $p < .05$ for women). The effect of projection on the number of ideas produced was not clear. However, the percentage of hostile ideas produced by those preferring projection was significantly greater than for those not preferring projection ($p < .05$ men; $p < .10$ women).

Blum (1954), Nelson (1955), and Eriksen and Browne (1956) have all provided evidence of a greater tendency to perceptual defensiveness on the part of those who prefer the "repressive" mechanisms of reversal and avoidance. Such findings might be related to the present theory in the following manner. It would be expected that if distances and differences between parts of a stimulus picture involving conflict were habitually neglected, the patterns and contours of such a picture would not be readily identified. On the other hand, any tendency toward picking out and emphasizing differences should make for easier identification. The "repressor" would in this case be more likely to show perceptual defense.

Such individual tendencies to exaggerate or to underestimate distance recall the interesting observations of Klein (1951) and his colleagues which have shown certain *Anschauungen* to characterize individual styles of adjustment. In particular, the tendency of some (levellers) to emphasize sameness and of others (sharpeners) to see differences is close to the present formulation at the descriptive level.

Generality of Styles of Dealing with Distance

It is important to emphasize once again that neither our theory nor our data imply that the style of reacting to distance is necessarily general over all situations for all people. As already noted it has been shown that defense preferences themselves are not completely general (Blum, 1957). Moreover, while the present study has revealed correlations between opposites measures and between opposites and defense preferences when conflict pictures were used to identify preferences, the correlations were not all equally high nor were they all significantly positive. A more systematic sampling of situations is required before the question about degree of generality of these or any other characteristics can be answered satisfactorily (Brunswick, 1947). Meantime, it appears that there is indeed some tendency toward generality, for a large proportion of the correlations between measures of opposites were positive and very few were negative. Furthermore, those preferring projection did more often give evidence of exaggerating distance in comparison to nonprojectors and those preferring reversal more often underestimated distance.

Nevertheless, a more qualified statement of the theory is required and one which recognizes some of the complexities of describing psychological structure. It may be stated in this fashion: if projection is chosen as a defense in some situation, then distances between the concepts and attitudes involved in the motivational system in question have a tendency to be exaggerated; if reversal is preferred, distances tend to be reduced. However, the exaggeration of intercategory distance by the projector may be accompanied by smaller mean differences *within* certain categories. In the present study the high projection–low reversal group showed significantly smaller mean differences between the attitudes toward the concepts constituting the ingroup category (FRIENDS, FAMILY, AMERICANS) than did the low projection–high reversal group. There was, however, no evidence of a DPI-related difference between the outgroup concepts here used.

Now, the observation that smaller ingroup attitude differences[15] and larger ingroup–outgroup differences characterized projectors need not embarrass the notion that certain manipulations of distance are characteristic of defensive styles, for a small mean ingroup difference may have the effect of increasing distance between ingroup and outgroup, the clarity of any such distinction between parts of a system resting as it does on the magnitude of differences between the components of any one category as well as on the distances between these components (ingroup) and the components of the other category (outgroup). Table 8 suggests that for projectors the mean intercategory distance[16] and the mean intracategory distance (ingroup) are more different (difference in means $= 36.0$; $s_d = 4.8$) than was the case for the low projection–high reversal group (difference in means $= 19.3$; $s_d = 4.3$).

---

[15] There is a question as to whether the small ingroup differences for high projectors may be due in some degree to their more favorable ingroup attitudes which might impose a ceiling effect which would tend to reduce the variation in ingroup attitudes.

In short, if for the reversal group distances between the ingroup components are not distinguishable from the distances that separate ingroup and outgroup components, boundaries will not be clear between ingroup and outgroup. Indeed, such an organization approximates one instead of two categories, and suggests that there has been assimilation of ingroup and outgroup attitudes making the former less favorable and the latter more favorable. On the other hand, contrast may be involved in the greater ingroup–outgroup attitude differences of projectors (Peak, 1958a).

SUMMARY

In a program concerned with psychological structure and attitude change it is necessary not only to find ways of discovering the existence of different structures but also to determine how they are reflected in attitudes and other psychological processes. Studies have been reported on the interrelation of measures of opposites structure, and the relation of such measures to preference for certain defense mechanisms, to the organization of ingroup and outgroup attitudes, and to other variables.

The psychological distance between events and the degree of their oppositeness was assumed to be a determinant of response in the following situations: (*a*) ratings on evaluative and potency scales of 19 concepts consisting of 6 pairs of opposite terms and 7 other terms, (*b*) the pattern of sorting statements about Negroes into categories, and (*c*) giving free associations to a number of words. Defense preferences were determined by the Blum Defense Preference Inventory. Attitudes toward the ingroup (friends, family, Americans), toward the outgroup (Negro roommate, Negro neighbors, foreigners), and

TABLE 8

WITHIN AND BETWEEN-CATEGORY DIFFERENCES IN ATTITUDES

	High Projection–Low Reversal Preference		*P* of Difference in Means	Low Projection–High Reversal Preference	
	Mean	*s*		Mean	*s*
|IG-OG|	46.1	24.4	<.05	34.4	20.5
|ME-OG|	33.4	24.1	<.10	24.8	20.1
|IG-ME|	17.1	13.4	>.10	20.1	12.1
|IG Diff|	10.1	8.9	<.05	14.9	11.2
N	35			37	

---

[16] The IG–OG difference entered in Table 8 is the mean of the mean of the nine differences between three ingroup and three outgroup attitudes calculated for each *S* and divided by three. This procedure makes the intracategory and the intercategory differences comparable.

toward the self were inferred from ratings of concepts on Semantic Differential scales. Approximately 100 students served as *S*s.

1. Preference for projection and for reaction formation or reversal were negatively correlated as had been predicted.

2. There was support for the prediction that high preference for projection and low preference for reversal would be related to a tendency to opposites structuring, and these relations were clearest when the defense preference was determined from the Blacky pictures which reflected conflict in an individual.

3. As predicted, greater distances between ingroup and outgroup attitudes were positively associated with a preference for projection, except when *S*s were high in preference for both reversal and projection. Projection was also related to favorable ingroup attitudes (significant in women), to unfavorable outgroup attitudes (significant in men), and to larger differences in ME–outgroup attitudes.

4. As predicted, high preference for reversal was positively related to favorable attitudes toward the outgroup in both men and women. There was also a significant negative relation of reversal scores with ingroup–outgroup distance, as well as with ME–outgroup distance.

5. It was also shown that those preferring reversal tended to produce fewer total ideas in writing the Blacky stories than did those not in that DPI category. This is also interpreted as depending on a style of dealing with distance.

6. It is proposed that these different correlates of projection and of reversal preference may be understood by a theory which assumes that when projection is adopted as a defense, distances between attitudes (and concepts) involved in a motive system are increased by contrast to the point where relations between the concepts (ingroup and outgroup) are disrupted and motivation reduced. On the other hand, when reversal is the preferred defense, motivating activity is removed defensively by treating differences between these events as though they were negligible, thereby producing assimilation and reducing disparity in the system below the amount that must be present in order for activity to continue.

7. Scores on the California F Scale had the expected positive relations to ingroup–outgroup attitude difference. High F scores were associated with high preference for projection but with low preference for reversal. This correlation with reversal contradicts the usual assumption of a positive relation between these variables.

8. As predicted, women tended to prefer reaction formation as a defense more often than did men and men preferred projection more often than women did. Women had consistently higher average opposites scores than men on every measure, except opposites associations. The different relations between different variables tended to be weaker for women than for men with the exception of relations to F scores.

REFERENCES

ADORNO, T. W., FRENKEL-BRUNSWICK, E., LEVINSON, D. J., & SANFORD, R. N. *The authoritarian personality.* New York: Harper, 1950.

BLUM, G. S. *The Blacky Pictures: A technique for the exploration of personality dynamics.* New York: Psychological Corp., 1950.

BLUM, G. S. An experimental reunion of psychoanalytic theory with perceptual vigilance and defense. *J. abn. soc. Psychol.,* 1954, **49,** 94–98.

BLUM, G. S. Defense preferences in four countries. *J. Proj. Tech.,* 1956, **20,** 33–41.

BLUM, G. S. Assessment measures. Unpublished manuscript, 1957.

BLUM, G. S., & HUNT, H. F. The validity of the Blacky Pictures. *Psychol. Bull.,* 1952, **49,** 238–250.

BRUNER, J. S., GOODNOW, J. J., & AUSTIN, G. A. *A study of thinking.* New York: Wiley, 1956.

BRUNSWICK, E. *Systematic and representative design of psychological experiments.* Berkeley: Univer. California Press, 1947.

DIXON, W. J., & MASSEY, F. J. *Statistical analysis.* New York: McGraw-Hill, 1951.

ERIKSEN, C. W., & BROWNE, C. T. An experimental and theoretical analysis of perceptual defense. *J. abn. soc. Psychol.,* 1956, **52,** 224–230.

GOLDSTEIN, S. A projective study of psychoanalytic mechanisms of defense. Unpublished doctoral dissertation, Univer. of Michigan, 1952.

HOVLAND, C. I., & SHERIF, M. Judgmental phenomena and scales of attitude measurement: Item displacement in Thurstone scales. *J. abn. soc. Psychol.,* 1952, **47,** 822–832.

KATZ, D. MCCLINTOCK, C., & SARNOFF, I. Measurement of ego-defense related to attitude change. *J. Pers.,* 1957, **25,** 465–474.

KLEIN, G. S. The personal world through perception. In R. R. Blake & G. V. Ramsey, (Eds.), *Perception; An approach to personality.* New York: Ronald, 1951.

KREEZER, G., & DALLENBACH, K. W. Learning the relation of opposition. *Amer. J. Psychol.,* 1929, **41,** 432–441.

NELSON, S. E. Psychosexual conflicts and defenses in visual perception. *J. abn. soc. Psychol.,* 1955, **51,** 427–433.

OSGOOD, C. E. Development and application of semantic differential. Unpublished manuscript, 1955.

OSGOOD, C. E., SUCI, G. E., & TANNENBAUM, P. H. *The measurement of meaning.* Urbana: Univer. Illinois Press, 1957.

PEAK, H. Attitude and motivation. In M. Jones (Ed.), *Nebraska symposium on motivation.* Lincoln: Nebraska Univer. Press, 1955.

PEAK, H. Psychological structure and psychological activity. *Psychol. Rev.,* 1958, **65,** 325–347. (a)

PEAK, H. Generalization of attitude change within an opposites structure. I. *USN tech. Rep., ONR,* Nonr 1224 (10) NR 171–039, June 1958, No. 3. (b)

PEAK, H. Generalization of attitude change within an opposites structure. II. *USN tech. Rep., ONR,* Nonr 1224 (10) NR 171–039, June 1958, No. 4. (c)

PEAK, H. Psychological structure and person perception. In R. Tagiuri & L. Petrullo (Eds.), *Person perception and interpersonal behavior.* Stanford: Stanford Univer. Press, 1958. (d)

PEAK, H. Attitudes, opposites structuring and F scores. *USN tech. Rep., ONR,* Nonr 1224 (10) NR 171–039, May 1959, No. 5. (a)

PEAK, H. Some correlations of opposites structure. *USN tech. Rep., ONR,* Nonr 1224 (10) NR 171–039, June 1959, No. 7. (b)

PETTIGREW, T. F. The measurement and correlates of category width as a cognitive variable. *J. Pers.,* 1958, **26,** 532–544.

RIESS, B. F. Genetic changes in semantic conditioning. *J. exper. Psychol.,* 1946, **36,** 143–152.

SHAPIRO, D., & TAGIURI, R. Sex differences in inferring personality traits. Unpublished ONR Report, 1959.

WOODWORTH, R. S. *Experimental psychology.* New York: Holt, 1938.

(Received August 10, 1959)

Vol. 74, No. 9 Whole No. 496, 1960

Psychological Monographs: General and Applied

THE EFFECTS OF EXPERIMENTALLY CONTROLLED EXPERIENCE UPON RECOGNITION RESPONSES[1]

ARNOLD BINDER AND SOLOMON E. FELDMAN

Indiana University

THE report presented in this monograph represents the accumulated evidence from the first four years of a continuing project of research in recognition. The emphasis in the series of experiments reported herein has been upon the effects of variations in the conditions of perceptual learning on the responses of recognition to ambiguous cues. In one or more of these experiments the learning factors of frequency, selective attention, reward, and punishment have been investigated in terms of their relationships to perceptual responding. The ambiguous cues to which the perceptual responses were made consisted of parts of the stimuli seen in previous learning trials.

There has been a good deal of interest in recent years in finding the relationship between certain aspects of past experience and various phenomena which are clearly perceptual or which are closely related to perception. A large proportion of the work in this area has involved studies with primary emphasis upon the effects of variations in experiential frequency upon perception. In some of these studies the past experience has meant the individual's personal history in his pre-experimental environment, while in others past history has implied certain experimentally controlled learning trials prior to some perceptual test. Thus, Howes and Solomon (1951) found that visual recognition threshold, as measured tachistoscopically in an ascending series, approximated a decreasing linear function of the logarithm of word frequency as tabulated by Thorndike and Lorge (1944). The obtained correlations between mean duration threshold and word frequency varied between −.68 and −.82, in separate determinations. The counts of Thorndike and Lorge were assumed to be estimates of the relative frequencies with which their subjects (*S*s) had encountered the various words. Attneave (1953) postulated that a high degree of familiarity with stimulus materials "may serve as a substitute for actual stimulation in the perceptual process" (p. 81). In his experiment *S*s were required to judge the relative frequency of occurrence, in English text, of each letter of the alphabet. He found that a power function provided a reasonably adequate fit for the points plotted on the basis of median judged frequency and actual frequency of occurrence of the letters in a sample of 10,341 letters from printed material. The correlation between log judged frequency and log actual frequency was .88.

Solomon and Postman (1952) emphasized that an individual *S*'s relative frequency of contact with units of English language is only crudely estimated by counts of printed material. Accordingly, in order to have better control of frequency, they experimentally manipulated the frequency with which *S*s used pronounceable nonsense words. Visual recognition threshold for words was reported as a negatively accelerated decay function of frequency of experimental exposure with zero frequency included in the analysis, and as approximately a decreasing linear function of the logarithm of frequency when the zero frequency was omitted. Arnoult (1956), more recently, exposed 10 random nonsense stimuli to *S*s at varying frequencies and then asked these *S*s to rate the same stimuli on a familiarity scale containing five points. He found rated familiarity of nonsense shapes a monotonic, negatively accelerated function of the experimentally controlled frequency of experience.

---

[1] This research was supported by Research Grants M-1259 and M-2170 from the National Institute of Mental Health of the National Institutes of Health, United States Public Health Service.

Besides the research on the effects of frequency on perceptual responses, there have been studies on the influence of such factors as deprivation, reward, punishment, set, meaning, avoidance training, and labeling upon perception. Thus, Sanford (1936) had *S*s give their interpretations of ambiguous pictures shortly before and shortly after a regular meal, and then counted the number of responses involving food and eating. He found many more food-related responses before as opposed to after the meal. Later, Sanford (1937) and Levine, Chein, and Murphy (1942) obtained results providing further confirmation of Sanford's hypothesis relating food deprivation and an increase of perceptual responses involving food. Proshansky and Murphy (1942) had *S*s estimate the lengths of intermediate lines, the weights of intermediate objects, and the extent of autokinetic movement, before and after long lines, heavy objects, and movement of a light to the right were rewarded while short lines, light objects, and movement of the light to the left were punished. Reward and punishment consisted, respectively, of giving 15 cents to and removing 15 cents from the *S*s. These *S*s showed a significant shift in their estimates in the direction of the rewarded stimuli.

Bruner, Miller, and Zimmerman (1955) found that errors of recognizing words, presented in conjunction with masking noises, decreased as knowledge of the set of possible alternatives increased. The effect was greater for short lists of words than for longer lists. Taylor (1958) gave *S*s preliminary training on nonsense syllables, such that one group saw only the syllables while the other group saw the syllables in association with pictures which presumably gave the syllables meaning. The tachistoscopic thresholds for the words were then determined, and she found that prior knowledge of the stimuli did lower thresholds, but meaning did not. Lazarus and McCleary (1951) conditioned the GSR to nonsense syllables by means of electric shock during a training period. Nonshocked syllables were also presented during this period. In later tachistoscopic trials they found the mean GSR for wrong responses higher when a shock syllable was shown than when a nonshock syllable was shown, indicating that *S*s can make "autonomic discriminations" even when they do not give accurate verbal reports. Finally, Hake and Eriksen (1956) found that the performance of *S*s on a recognition task was affected by the number of labels used in naming the actual patterns used in the task, but not by the number of labels used in naming the same patterns previously.

The experiments to be reported here fall in the general class of research aimed at determining the relationships between experiential and perceptual variables. The experiential variables in these experiments are frequency, reward, and punishment; while the perceptual variables involve responses indicating recognition of ambiguous stimuli. An experimental approach to evaluate and control the experience has been used.

Recognition has been assumed to be a classifying process where cues are used to make a set of decisions as to the assignment of a given stimulus object to a particular class. Binder (1955) and Bruner (1957) have adopted this approach in their theoretical presentations of the processes of recognition and perception, respectively. According to this viewpoint, a naming response designates a particular stimulus object as belonging to a class defined on the basis of common cues. Thus, when a person is confronted with the task of identifying an orange, he may use its size to rule out such response classes as "basketball" and "beach ball," and perhaps its shape to rule out "tangerine." One would predict, on the basis of this analysis, that, when an individual has learned to make discriminatory responses indicating recognition on the basis of configurations of cues, he can also identify objects on the basis of subsets of the configurations.

But, if the complete set of cues is necessary for unique discrimination, any reduction in information resulting from the availability of only a subset of all cues produces ambiguity. If a person catches only an instantaneous glimpse of a spherical object traveling through the air at high velocity, for example, we assume that his previous experience with full cue objects enables him to use the reduced cues for identification purposes. The cue "spherical," however, has been part of the complete set of cues of many objects, including tennis balls, golf balls, and oranges. The situation is ambiguous because of the limited information in the form of cues.

Ambiguity may also occur when too many cues are present, as, for example, in the case of reversible figures where the cues defining two or more objects are present in the stimulating situation. Thus, considering stimulus–object classes as defined on the basis of common attributes or cues, a stimulus situation may be called ambiguous when the entire set of cues available to an observer belongs to the attribute arrays of two or more such classes.

The tachistoscope produces ambiguity by presenting a stimulus object so rapidly that details are obliterated in the over-all blur seen by an observer. The cues available to the observer might have come from any number of different objects. And the faster the stimulus is presented, the more the blurring effect and the greater the ambiguity as a result of the larger number of stimulus objects from which the available cues could have come.

In the case of the Rorschach and similar projective techniques, ambiguous stimuli are used to assess various aspects of the life histories of individuals. An unstructured form, which has shape, color, and/or texture-like attributes common to many object classes, is presented to *S*s under instructions to specify what it looks like. Some may call a given blur a human, others a bear, and perhaps still others a dead eagle. Assuming no pathology on the part of the *S*, attributes of the blur may be characteristic of the objects in each of these classes. The attributes which would lead to clear differentiation among the classes are not available.

But what are the factors which lead to one response rather than another equally adequate one in an ambiguous situation? Which response, for example, would the person, who caught the quick glimpse of the object in the air, make if asked to identify the object. If he were standing on a tennis court, he would almost certainly call it a "tennis ball," unless, perhaps, there were an adjoining golf driving range. The response would therefore clearly be related to the frequency with which the various objects were previously experienced in the given context. Using the Rorschach and similar projective techniques, the clinician uses the responses of *S* to evaluate the latter's personality. The clinician is, explicitly or implicitly, assuming that the choice among the possible recognition responses of an individual to an ambiguous blot configuration is a function of the individual's previous experiences of reward, punishment, deprivation, frustration, and so forth. One person may have been frustrated in his experiences with women and give no responses dealing with femininity throughout his entire Rorschach protocol. Another may have been deprived sexually and report sexual symbols on every blot.

Much of the research reported above on the relationship between experience and perception has dealt specifically with the factors which influence responses to ambiguous cues (Levin et al., 1942; Proshansky & Murphy, 1942; Sanford, 1936, 1937; Solomon & Postman, 1952). But there has been no systematic attack on the problem, and wide gaps remain in our knowledge. The purposes of the present experimental program were to investigate certain relationships not considered in the earlier work and to explore further some of the factors previously evaluated. Particular emphasis was laid upon the analysis of the effects of frequency of occurrence upon responses to ambiguous cues.

EXPERIMENT I

There were two principal goals to this experiment. The first was to investigate the functional relationship between the frequencies with which stimulus figures are experimentally experienced and the frequencies of responses to cues which are parts of two or more of these figures. At one point in the presentation of his model of recognition, Binder (1955) assumed that an "individual uses a frequency of designation of a particular class name equal to $P_i n$ where P_i is the objectively measured probability of occurrence of the given class and n is the number of repetitions of the set of conditions" (p. 123). If we consider the probability of occurrence of an object equal to the relative frequency of its previous occurrence in the given context, this leads to the prediction that the relative frequency of a particular response to ambiguous cues is equal to the relative frequency of occurrence of that response and its associated stimulus. The second goal was to test the hypothesis that *S*s who learn to discriminate among objects on the basis of configurations of cues learn to make accurate classifying responses on the basis of parts of these configurations. As stated above, this is an implicit assumption in the models of Binder (1955) and of Bruner (1957).

Method

Subjects. The *S*s were all volunteers from various sections of introductory psychology at Indiana University. The learning task was very difficult, and only *S*s who were prepared to spend up to two hours in the experiment were used. The data of about one-fourth of the participating *S*s had to be eliminated because of administrative errors, the failure of some to reach the first criterion level by about 90 min., and the necessity for some to leave near the 2-hr. mark. The average time needed for completion of the experiment was about 85 min. All *S*s were run individually, but grouped according to procedure. The data of 35 *S*s in Group I-E and 31 in Group I-U were usable.

Apparatus. The *S*s sat about 5 ft. from the exposure apparatus which consisted of a plywood panel with two separated 8 in. high by 8¼ in. wide one-way vision mirrors located in approximately the center of the panel. Attached to the plywood panel, on the experimenter's (*E*'s) side, was a box with an open front face and grooved sides to hold the stimulus and response cards. The cards were illuminated by four (for each mirror) six-watt bulbs which were placed behind the mirror so as to provide relatively even illumination of the surfaces of the cards and yet not be visible to the *S*s from their normal viewing position. Panel, table, and box were all painted a dull black.

The apparatus was situated in a semidarkened room, and *S*s could see the stimulus and response cards only when the bulbs behind the mirrors were on. At other times, of course, they could only see the reflecting surfaces of the mirrors. The mirrors were designed to provide 33% transmission of light, 33% reflection, and absorption of the remaining light.

Presentation materials. Eight nonsense syllables of 33% association value were used as names during the learning trials of the experiment (Glaze, 1928). The names are as follows: TUZ, RUX, WOY, JOM, KAL, NAF, VED, and GEP. Eight stimulus figures were also used and may be seen in Fig. 1. The squares within which the differentiating cues are placed are 4 in. in length on each side. These stimulus figures were designed in such a way that every cue within the common square was necessary for some discrimination.

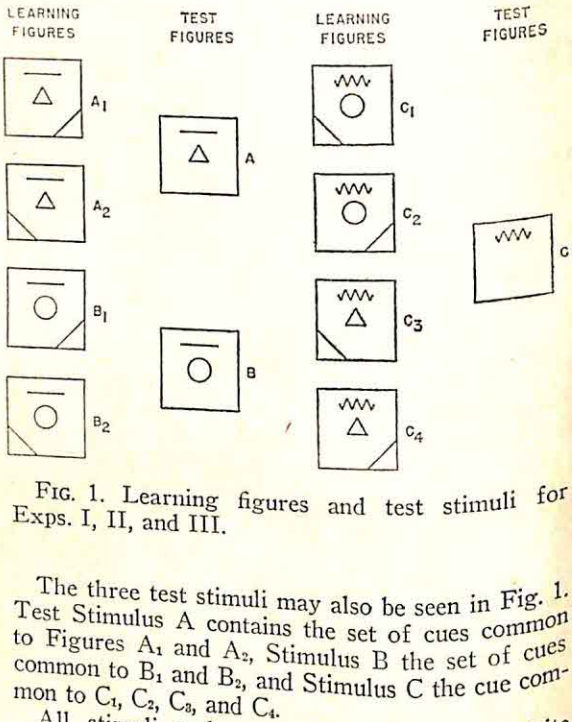


FIG. 1. Learning figures and test stimuli for Exps. I, II, and III.

The three test stimuli may also be seen in Fig. 1. Test Stimulus A contains the set of cues common to Figures A_1 and A_2, Stimulus B the set of cues common to B_1 and B_2, and Stimulus C the cue common to C_1, C_2, C_3, and C_4.

All stimuli and names were drawn on white drawing board. Great care was taken during the drafting to ensure that each cue common to four stimulus figures, such as the circle or the straight line at the upper part of the square, appeared in the same location of the exposure surface when the respective figures were placed in the grooves on the box. Similarly, the cues appearing on the test cards were placed so as to provide for alignment on the viewing surface of common cues on learning and test figures.

Procedure. All *S*s were trained by the method of paired associates (with correction) during the learning phase, with stimulus figures appearing in the left mirror and nonsense syllables in the right mirror. The following intervals were used: stimulus exposure 5½ secs., response exposure 2 secs., intertrial interval 7½ secs., and delay after warning buzzer 1½ secs. The delay between blocks averaged about 30 secs., but was never less than 22 secs.

The procedures used for the *S*s in Group I-E and those in Group I-U differed only in the composition of each learning trial block. For the *S*s in Group I-E, a trial block consisted of a single presentation of each of the eight stimulus figures together with its associated nonsense name. The trial blocks for the *S*s in Group I-U contained figure-name pairs presented varying numbers of times. In the trial blocks of Group I-U, either A_1 or A_2 occurred twice and the other once, B_1 or B_2 occurred four times and the other once, and one of C_1, C_2, C_3, or C_4 occurred three times and the other three once each. While A_1 and A_2 always appeared in the ratio of two to one per trial block, the particular one which occurred twice and the one which occurred once were randomly determined for different *S*s. In a similar fashion, whether B_1 or B_2 appeared four times per block and whether C_1, C_2, C_3, or C_4 appeared three times per block were randomly determined. Table 1 shows a summary comparison of the trial blocks for the two groups. This table can be considered only representative, of course, since the particular one of a subset (like A_1 or A_2) which occurred more frequently varied randomly over *S*s. The Group I-U trial blocks in all cases consisted of 14 pairings.

The nonsense syllables paired with particular figures were randomly determined for different *S*s. The *S*s were instructed that the associated nonsense syllables were to be considered the names of the figures and that they were to anticipate each figure's name after the figure was exposed but before its nonsense name appeared. To make this more like a recognition situation, the *S*s pronounced rather than spelled out the nonsense syllables. During the first trial block, *E* pronounced each syllable as it appeared, in order to minimize any reluctance on the part of *S*s to pronounce any syllable.

TABLE 1

Number of Occurrences of Learning Figures per Trial Block for One Possible Assignment in Exp. I

Learning Figure	Group I–E	Group I–U
A_1	1	2
A_2	1	1
B_1	1	4
B_2	1	1
C_1	1	3
C_2	1	1
C_3	1	1
C_4	1	1
Total Presentations per Block	8	14

The ordering for the presentation of stimuli for each odd-numbered trial block was determined by a randomization schedule which *E* had prepared in advance. A restriction imposed on the randomized order was that a single stimulus could not appear more than twice in succession. The reverse order of presentation to that of the immediately preceding odd-numbered block was used for each even-numbered block. The sequence of randomized order as read from the schedule, reversal of that order, new randomized order, etc. was repeated with only one break until a criterion of two consecutive blocks of perfect anticipations was reached. This break consisted of a 5-min. rest period at the end of 40 min. (if the criterion had not been reached by this time). At the criterion point of two perfect blocks, the *S*s were given another 5-min. rest period. After this rest period, the presentation of learning blocks of paired associates was repeated until the *S*s achieved one more block of perfect anticipations. They were then given their first test trial, and this was followed by the learning trials necessary for another block of perfect anticipations. Then came the second test trial, learning trials to one further block of perfect anticipations, and finally the last test trial. The purpose of the learning blocks intervening between all test trials was to assure as far as possible that the *S*s maintained the learning level achieved previously.

The three test trials which all *S*s were given consisted of single presentations of each of the Stimuli A, B, and C (see Fig. 1). The order of presentation of these three stimuli was randomized with the one restriction that Stimulus C was not to appear on the first trial. The test stimuli were shown for a duration of 1¼ secs., beginning 1½ secs. after the sounding of the warning buzzer. The intertrial interval was 9 secs. For the test phase of the experiment only the left mirror of the exposure apparatus was used.

Immediately following the first criterion of two consecutive perfectly anticipated trial blocks, the *S*s were told that they were ready for the second phase of the experiment in which a new figure would be presented after each further block of all correct responses. They were told that the new figure would not be exactly like any of those shown before and would be exposed for a shorter period of time, but that they were to respond with the name of the most closely related figure. The additional instructions during the actual test phase, after the 5-min. break, were as follows:

> In order to make sure that you still remember the characteristics together with the correct name of each of the figures presented during the first part of the experiment, I will now present them in the same way once again. As before,

I would like you to name the figure shown in the left window just as soon as you can after it is exposed, before the name is shown a few seconds later in the right window. We will not go on to the new figure, which I told you about before, until you complete a run giving the correct name of all of the figures before their names are shown in the right window. Correct any misses or mistakes as you did previously.

[After criterion of one perfect block]:

Now here is a new figure. Give the one name out of the set you have learned which seems most appropriate, even though this figure is not exactly like any of those just shown. No name will appear in the right window. (The figure became visible to the *S*s about 9 secs. after the completion of these instructions.)

[After exposure of first test stimulus]:

Once again we will rerun the figures and names as in the first part of the experiment to make sure you remember them. Name the figure before its name appears in the right window and correct misses and errors.

After the criterion of one further errorless block was achieved, the same instructions were read as after the perfect block above ("Now here is a new figure," etc.). These were repeated once again before the final test trial. The instructions which followed the first test stimulus ("Once again we will rerun," etc.) were reread prior to the final set of trials of paired associates.

Results and Discussion

Let us consider an example in terms of the responses to Test Stimulus A (see Fig. 1) in order to illustrate how the results may be analyzed. If an *S*'s response to A is the name of either Figure A_1 or Figure A_2, it is called an appropriate response. If he responds with one of the names of any of the other six figures, it is called an inappropriate response. A third category of response could occur if an *S* did not follow instructions and responded to a test stimulus with a name which had not occurred during learning. This happened in very few cases, and these responses were placed in the inappropriate category. Thus, we can consider the results in terms of the number of appropriate responses to A as opposed to the number of inappropriate responses, and in terms of the relative number of A_1 and A_2 naming responses to A. Similar analyses may be made of the responses to Test Stimuli B and C.

The numbers of appropriate and inappropriate responses as well as the numbers of responses of differing learning frequency within the appropriate category may be seen in Tables 2 and 3. The frequencies expected from the hypothesis that relative frequency of response equals relative frequency of learning occurrence are shown parenthetically in the columns designated "Appropriate Responses." The correspondence between the values expected on the basis of this hypothesis and those actually obtained is obviously close. However, it will be most convenient to delay a fuller discussion of these particular results until the data are analyzed in terms of the appropriate–inappropriate dichotomy.

According to the second hypothesis stated in the introduction to Exp. I, *S*s who had learned the figure–nonsense name associations perfectly might have been expected to have made no inappropriate test responses. But, despite the fact that a strin-

TABLE 2

OBTAINED AND EXPECTED (IN PARENTHESES) RESPONSE FREQUENCIES TO TEST STIMULI FOR *S*s IN GROUP I–E

Test Stimulus A				
Appropriate Responses				Inappropriate Responses
A_1	A_2			
15 (15.0)	15 (15.0)			5
Test Stimulus B				
Appropriate Responses				Inappropriate Responses
B_1	B_2			
15 (14.0)	13 (14.0)			7
Test Stimulus C				
Appropriate Responses				Inappropriate Responses
C_1	C_2	C_3	C_4	
8 (8.25)	7 (8.25)	7 (8.25)	11 (8.25)	2

TABLE 3

OBTAINED AND EXPECTED (IN PARENTHESES) RESPONSE FREQUENCIES TO TEST STIMULI FOR *S*S IN GROUP I–U

Test Stimulus A		
Appropriate Responses		Inappropriate Responses
A(2)	A(1)	8
20	3	
(15.3)	(7.7)	

Test Stimulus B		
Appropriate Responses		Inappropriate Responses
B(4)	B(1)	5
20	6	
(20.8)	(5.2)	

Test Stimulus C		
Appropriate Responses		Inappropriate Responses
C(3)	All C(1) Pooled	3
13	15	
(14.0)	(14.0)	

gent criterion was used, *S*s made errors in terms of misclassification up to the final one of the required five perfect blocks. A factor based on incorrect anticipations during the learning trials following the achievement of the first three required perfect trial blocks was accordingly used to estimate the imperfection in learning at criterion.

In order to make a correct anticipation during learning, a given *S* must use all the information in the form of cues: circle or triangle, wavy or straight line at the top, and line at right or left corner (see Fig. 1). He must, in other words, classify correctly according to three dichotomies. An incorrect response may result from an error in one, two, or three of those dichotomous categorizations. If the figure exposed happens to be B_2, for example, a response consisting of the name of B_1 or A_2 would result from one misclassification while the name of C_4 to B_2 would represent three misclassifications.

The computations used in arriving at the expected number of appropriate and inappropriate test responses on the basis of the errors in anticipation made by *S*s at criterion are outlined in Table 4. For all *S*s, total number of misclassifications on all trials, following the first three required perfect trial blocks, were counted. Also counted were the total number of blocks, over all *S*s in the respective groups, after the third perfect one and up to and including the fifth perfect one. Since each block for Group I-E contained eight trial pairings and three classifications were necessary per trial (one for each cue), the total number of necessary categorizations for the *S*s in Group I-E after the third perfect block is the total number of trial blocks multiplied by 24 (8 × 3). The multiplier in the case of the *S*s in Group I-U is 42 (14 × 3) since there were 14 trials per block for Group I-U. The number of misclassifications per attempted categorization at criterion for each group was obtained by dividing total misclassifications by the number of necessary categorizations.

The number of inappropriate responses expected purely on the basis that the *S*s had not learned to classify without error using each of the three cues could now be computed. When presented with Test Stimuli A or B (see Fig. 1) each *S* had to categorize on the basis of two cues, while when presented with C he had to make only one dichotomous categorization (one of the C_i [$i = 1, 2, 3, 4$] stimulus figures or not a C_i stimulus figure). The total number of categorizations necessary by the *S*s in Group I-E to Stimulus A was 70 (35 × 2); to B, 70; and to C, 35. The comparable numbers for the *S*s in Group I-U were 62, 62, and 31. The total number of necessary test

TABLE 4

PERFORMANCE IN LEARNING TRIALS FOLLOWING FIRST THREE REQUIRED PERFECT BLOCKS IN EXP. I (Summed Over All *S*s)

	Group I–E	Group I–U
Total Misclassifications	91	83
Trial Blocks	128	113
Necessary Categorizations	3072	4746
Misclassifications per Attempted Categorization	.0296	.0175

TABLE 5

Frequencies of Appropriate and Inappropriate Responses to Test Stimuli for Both Groups of *S*s in Exp. I

Group	Test Stimulus	Obtained Responses		Expected Responses	
		Appropriate	Inappropriate	Appropriate	Inappropriate
I–E	A	30	5	33	2
I–E	B	28	7	33	2
I–E	C	33	2	34	1
I–U	A	23	8	30	1
I–U	B	26	5	30	1
I–U	C	28	3	30	1

categorizations multiplied by the number of misclassifications per categorization at criterion gives the total expected inappropriate responses to the test stimuli.[2]

The expected frequencies of appropriate and inappropriate responses and also the obtained frequencies may be seen in Table 5. Fisher's exact test was applied to all differences between expected (on the basis of actual learning responses) and obtained frequencies, and only the difference due to the responses of the *S*s in Group I-U to Test Stimulus A is significant. The significance level is $P \leqq .05$. It would seem reasonable to conclude that these results support the hypothesis that *S*s can identify components of the stimuli to which they previously learned names.

Perhaps a final remark should be made on the justification for choosing the learning trial blocks following the third required perfect block for predicting frequencies of inappropriate test responses. Each *S* had to achieve five blocks with perfect anticipations: two consecutive ones during the learning phase, one more after a 5-min. rest period, and one additional perfect block after each of the succeeding two test trials. It was necessary to be sure that *S*s were at criterion level of performance for these computations of misclassifications and the trial blocks prior to the first two perfect ones could obviously not be used. Nor could the trials between the second and third perfect blocks be used since a rest period, and presumably some forgetting, had intervened between these blocks. After the third perfect block, however, only a 1¼-sec. test trial and a very brief set of instructions came between successive learning trial blocks. This made it reasonable to assume that the *S*s were not below their first criterion level during this period.

For the sake of facilitating future discussion (with specific reference to the *S*s in Group I-U) it is convenient to introduce notation to allow us to refer to the stimulus figure occurring at a given frequency per block, regardless of the specific figure it may be. Accordingly, let A(1) be the A_i $(i = 1$ or $2)$ figure which occurs once per block for a certain *S* (as randomly determined) and A(2) the A_i figure occurring twice per block for the *S*. Thus if A_1 occurs twice per block and A_2 once as a result of the random determination, A(1) refers to A_2 and A(2) to A_1 in this particular case. In a similar fashion, let B(1) be the B_i $(i = 1$ or $2)$ occurring once per block, B(4) the B_i occurring four times per block, C(1) any of the three C_j $(j =$ 1 or 2 or 3 or 4$)$ figures occurring once per block, and C(3) the C_j figure occurring three times per block for a given *S*.

Let us return to the comparative frequencies which were obtained within the category of appropriate response. Table 2 contains the frequencies of the responses of the *S*s in Group I-E to the three test stimuli. The numbers in parentheses were obtained for a given test stimulus by the formula

$$\frac{O_i}{O_s}(N)$$

where O_i represents the number of times the i^{th} figure occurred on each learning trial block, O_s represents the number of times the learning figures containing all cues of the given test stimulus occurred per trial block, and N represents the total number of appropriate responses given to the test stimulus by all *S*s. As an example of the use of this formula, consider the responses to Test Stimulus A (Table 2). In this case the total number of appropriate responses, or N, is 30. If the learning figure of interest (that is, the i^{th} figure) were A_2, $O_i = 1$. Since there are two learning figures containing all cues of Test Stimulus A (that is, A_1 and A_2) and they occurred a total of twice per trial block, $O_s = 2$. Therefore, we have

---

[2] This requires the assumption that each *S* who gives an inappropriate response makes only one misclassification in doing so. But this would seem entirely reasonable since the average number of misclassifications for each erroneous anticipation during learning was 1.09 for the *S*s in Group I and 1.02 for those in Group II. And to make the argument even stronger, the maximum number of misclassifications possible was three during the learning trials and only two during the testing trials.

$$\frac{O_i}{O_s}(N) = \tfrac{1}{2}(30) = 15$$

We thus predict a total of 15 responses to Test Stimulus A consisting of the name previously associated with A_2. The parenthetical numbers in Table 2 are in all cases, then, the frequencies of various appropriate responses which would be expected if the *S*s were giving particular appropriate responses at the same relative frequencies as the occurrence of their associated stimuli per learning trial block.

Table 3 contains the comparable obtained and expected frequencies for *S*s in Group I-U. The numbers in parentheses for a given test stimulus are not equal (as in Table 2) since the learning stimuli were presented at differing frequencies per trial block.

The relative frequencies expected on the basis of learning occurrences were considered theoretical (population) values in testing for the significance of the difference between obtained and expected relative frequencies. The binomial test was used for the five cases in which the appropriate responses fell in two classes, and the χ^2 test for the one case involving four appropriate classes of response (that is, the responses to Stimulus C for the *S*s in Group I-E). Only one of these six comparisons showed a significant difference ($P \leqq .05$). This again occurred in the responses to Stimulus A for the *S*s in Group I-U.

Since much of the above is consistent with recency, it was desirable to check upon whether or not the particular appropriate responses given tended to be the last encountered in learning. Accordingly, the learning block preceding each test trial for the *S*s in Group I-E was examined to see if the test response was or was not the most recent of the possible appropriate responses. It should be emphasized that, for this purpose, only appropriate test responses to a given test stimulus were considered and the sequence of figure–name pairs during the immediately preceding learning block was evaluated only from the standpoint of the ordering of appropriate responses. Thus the two appropriate responses to A are the names of A_1 and A_2 and, when either of these names was given, the prior learning block was examined to see if the stimulus–response pair corresponding to that name did or did not occur after the other possible pair which could have been named. The test responses to B and C were similarly analyzed. Since there are only two appropriate responses to Stimuli A and B, chance occurrence with no recency effect would be indicated by an equal division of the appropriate responses into "last" and "not last" categories. In the case of Stimulus C there are four appropriate responses, so one in four can be expected to be a last response by chance. Table 6 contains the number of appropriate responses to each of the three test stimuli which were and which were not the names of learning figures oc-

TABLE 6

NUMBER OF APPROPRIATE RESPONSES FOR GROUP I–E WHICH WERE AND WERE NOT LAST LEARNING RESPONSE (AS COMPARED WITH OTHER POSSIBLE APPROPRIATE RESPONSES)

	Last		Not Last	
Test Stimulus	Obtained	Expected by Chance	Obtained	Expected by Chance
A	12	15	18	15
B	14	14	14	14
C	10	8.25	23	24.75

curring last (ignoring inappropriate stimulus names) in the immediately preceding learning block. Table 6 also gives chance expectations for "last" and "not last" responses. Adding up all "last" and "not last" responses gives sums of 36 and 55, respectively. When one considers that purely random responding with regard to "lastness" gives expectations of 37.25 "last" occurrences and 53.75 "not last" occurrences, it is clear that the data present no evidence for a recency effect large enough to effect the above interpretations. However, it should be pointed out that this test is surely not sensitive enough to detect the small sequence effects predicted and found by such investigators as Estes and Straughan (1954).

As a final recency comparison, it was found that 16 of the total of 105 responses (appropriate and inappropriate) given by all *S*s to all test stimuli were the names of the very last stimulus (considering all eight learning figures in this case) to occur in the previous learning blocks. This compares with an expectation of 13.1 if the *S*s were giving the most recent of all learning responses purely on a chance basis.

Thus, a very striking relationship between the perceptual variable, frequency of response to an ambiguous testing stimulus, and the learning variable, frequency of occurrence of figures in blocks of paired associates, was found. The evidence supports the preliminary prediction that the frequencies of responses consisting of the various possible appropriate figure names parallel the frequencies with which the *S*s had previously responded to the figures.

In only one of the comparisons did this relationship fail to hold: that was in the case of the responses of the *S*s in Group I-U to Test Stimulus A. There were other idiosyncrasies in this particular case. First, as pointed out above, this was the only group that gave significantly more inappropriate responses than expected on the basis of misclassifications during learning trials at criterion. And second, if one looks at the particular kinds of inappropriate responses made, one finds that six of the eight errors consisted of the name of Figure C_4 to Test Stimulus A (see Fig. 1). Notice that C_4 differs from the appropriate response to A, A_1, only in that the former has a wavy line and the latter a straight line at the top. In all other test trials the inappropriate responses were much more evenly scattered over the possible choices. The five inappropriate responses of the *S*s in Group I-E to Test Stimulus A, for example, consisted of one C_2 name, one B_1 name, one B_2 name, and two C_4 names.

While it is hardly startling, support was also found for the hypothesis that *S*s who have learned differentiating responses to configurations of cues are able to make accurate discriminatory classifications on the basis of cue components. As mentioned above, the only discrepancy in six comparisons was the generally pathological case of the responses to Stimulus A by the *S*s in Group I-U.

EXPERIMENT II

The primary purpose of this experiment was to test further the hypothesis that the relative frequency of an appropriate recognition response to ambiguous cues is equal to the relative frequency of that response and its associated stimulus during previously reinforced learning trials. A secondary purpose was to explore the generality of this relationship by altering the conditions in the one previous experimental situation in which it did not hold. This alteration was accomplished by dividing the eight stimulus figures of Fig. 1 into two sets and then administering them to separate groups of *S*s. The division of figures was such that the appropriate response figures for the test stimulus giving the discrepancy in the previous experiment (that is, A) were in one part and the figure associated with the inappropriate response which seemed to be the source of difficulty (that is, C_4) was in the other. This division had the further very marked advantage of reducing the average time necessary to complete the experiment from the approximately 85 min. required of the *S*s in Exp. I to about 25 min. This facilitated the accumulation of much larger sample sizes, which seemed very desirable in attempting to verify the relationships found previously.

Method

Subjects. The *S*s were volunteers from the introductory psychology course at Indiana University. All *S*s were run individually but were considered to be in one of two groups in accord with the learning procedure used. Group II-AB contained 55 *S*s; and Group II-C, 54 *S*s. In contrast to the losses of Exp. I, very few *S*s were lost because of administrative difficulties in this experiment.

Apparatus and materials. The same apparatus, stimulus materials, and response words as in Exp. I were used.

Procedure. All *S*s were trained by the method of paired associates in the manner described for Exp. I, with only those changes necessitated by the different stimulus groupings. The stimuli used for the *S*s in Group II-AB were A_1, A_2, B_1 and B_2 during the learning trials and A and B during the test trials. The name associated with a given figure was randomly determined for each *S* from among the eight possibilities.

Each trial block for *S*s in Group II-AB was eight stimulus–response presentations in length; two of the four figures occurred more than once per block. The specific figures replicated varied randomly over *S*s. For some *S*s, Figure A_1 and its associated nonsense syllable occurred twice per block and Figure A_2 once, while for others A_1 occurred once and A_2 twice. Similarly, Figure B_1 occurred four times per block and B_2 once for some *S*s, while B_1 occurred once and B_2 four times for others. Once chosen for a given *S*, the assigned frequencies for the figures were maintained through all of his learning blocks. The following is an example of stimulus composition and ordering for a Group II-AB block: B_1, A_2, B_2, B_2, A_1, B_2, A_2, B_2.

The stimuli used for the *S*s in Group II-C were C_1, C_2, C_3, and C_4 during the learning trials and C during the test trial. Again four of the syllables were randomly matched with the learning figures for different *S*s. The learning trial blocks for the *S*s in Group II-C were only six figure–name pairings in length. One among C_1, C_2, C_3, and C_4 was exposed three times per trial block and the other three once each, but the more frequently occurring one was chosen randomly for different *S*s. Once chosen, this remained the figure which occurred three times through all the trial blocks for an *S*. A possible figure composition and ordering for a block of Group C is C_4, C_1, C_4, C_4, C_3, C_2.

Prior to the learning trials, the *S*s in both groups were told that the nonsense syllables were to be considered the names of the figures and that their task in subsequent trials was to anticipate this name after the figure was exposed but before the name appeared. The order of presentation within a block during learning was random on odd-numbered trial blocks (with the restriction of no more than two consecutive appearances) and the reverse of the previous odd-numbered block on even blocks. Learning proceeded to a criterion of two consecutive trial blocks with perfect anticipation. After reaching this criterion, all *S*s in both groups were given a 5-min. rest period, at the end of which the paired-associates blocks were repeated until they achieved one further block of perfect anticipations.

The Group II-AB *S*s were then given their first test trial, and this was followed by a further set of learning blocks to a criterion of one more perfect block. Then came the final test trial. During the two test trials of the *S*s in Group II-AB, Test Stimulus A was presented once and Test Stimulus B once in random order over *S*s (see Fig. 1).

Following the rest period and succeeding trials to one perfect block, the *S*s in Group C were given a single test trial consisting of Test Stimulus C. This terminated the experiment for them.

Prior to each test trial all *S*s were told that a new figure, somewhat different from all those used during learning, would be presented and that they were to respond with the one previously learned name which seemed most appropriate. They were informed of the criterion requirements in terms of necessary perfect trial blocks preceding all learning stages.

Results and Discussion

Table 7 contains the number of *S*s in Group II-AB who gave appropriate and inappropriate responses to Test Stimuli A and B. It also contains the number of *S*s who gave the various appropriate responses to Stimulus C in Group II-C. To illustrate the table's meaning, when presented with Test Stimulus A, 36 of the 55 *S*s in Group II-AB gave an appropriate response consisting of the name of the learning figure which occurred twice as frequently as the other in their learning blocks, that is, A(2); 16 *S*s made A(1) and three *S*s made inappropriate responses (B's or C's). The numbers in parentheses in Table 7 are the response frequencies we would expect if the *S*s responded with appropriate responses at the same relative frequencies as their occurrence in the learning trial blocks; they were obtained as described in Exp. I.

The very few inappropriate responses obtained makes it reasonable to account for the phenomenon of inappropriate responding solely on the basis of imperfect learning of name pairs. This confirms the previous finding.

TABLE 7

Obtained and Expected (in Parentheses) Response Frequencies to Test Stimuli for All *S*s in Exp. II

Test Stimulus A (Group II–AB *S*s)		
Appropriate Responses		Inappropriate Responses
A(2)	A(1)	B_1 or B_2
36	16	3
(34.7)	(17.3)	
Test Stimulus B (Group II–AB *S*s)		
Appropriate Responses		Inappropriate Responses
B(4)	B(1)	A_1 or A_2
42	11	2
(42.4)	(10.6)	
Test Stimulus C (Group II–C *S*s)		
C(3)	All C(1) Pooled	
37	17	
(27.0)	(27.0)	

The fit between obtained and expected frequencies for the responses of the *S*s in Group II-AB to Test Stimulus A, on the one hand, and Stimulus B, on the other, is remarkably close. Statistically, the small differences between obtained and expected values are very far indeed from approaching significance. These data surely lend support to the hypothesis that the relative frequencies of various recognition responses to ambiguous cues equal the relative frequencies with which these responses occurred previously in association with their stimuli.

However, using a χ^2 test, the difference between the obtained and expected responses to Test Stimulus C by the *S*s in Group II-C is significant at the .05 level. In looking at these data in Table 7, it should be remembered that the expectation of equal frequencies in the two appropriate response categories stems from the fact that, although one learning figure among the C_j occurred three times as frequently per learning block as each of the other three—that is, C(3)—all responses consisting of the names of the three less frequent learning figures—all C(1)—have been pooled. In effect this means that one expects as many responses consisting of the more frequent name as all three less frequent names combined.

These results are of particular significance when compared with the findings of Exp. I. In Exp. II, the *S*s in Group II-AB, who were tested on A after learning trials involving the same ratio of presentation of A_1 and A_2 as the *S*s in Group I-U (that is, 2 to 1), never saw the C_j learning figures. Consequently, the source of difficulty involving a large number of inappropriate C_4 responses to A was eliminated. The present data, therefore, support the hypothesis that the large number of inappropriate responses (of a particular kind) contributed toward producing the single significant discrepancy between relative frequencies during learning and relative frequencies among appropriate test responses in Exp. I.

The present data from *S*s in Group II-C are at variance with the previous finding of no significant difference between expected and observed frequencies for the same test stimulus and the same learning presentation frequencies. In fact, the obtained and expected values for Group I-U are almost identical. How can we account for this marked difference in results? Assuming that the difference was not due simply to chance sampling fluctuations, we must look at the essential difference between the two experiments and consider its possible effects. The *S*s of Group I-U were trained on all eight learning figures and tested on all three test stimuli, while the *S*s in group II-C were trained on only the four C_j learning figures and tested on C. The entire set of learning and test stimuli of the present experiment was only part of a larger set in the earlier experiment. In other ways—relative frequency of presentation among learning figures, learning criterion requirements prior to testing, nonsense syllables potentially paired with learning figures, over-all method of learning and test presentation, and instructions—the two experiments were identical.

In the case of the *S*s in Group II-AB as well as in the case of the *S*s in Group I-U, the trial blocks for all *S*s contained at least two stimuli which appeared more than once. This was not the case with the *S*s in Group II-C who had only one stimulus presented more than once per block. When in repeated blocks of four different stimuli, one stimulus appears three times and the others appear only once each in successive blocks of six trials, it seems reasonable to assume that the more frequent stimulus becomes relatively conspicuous. Thus, 49% of the *S*s made no errors in identifying the C(3) stimulus throughout their learning trial blocks (excluding the first block, of course, during which *E* read aloud the names). Moreover, 59% made no errors in identifying C(3) during either the second, third, or fourth blocks; and 66% made no errors in identifying it during the second block alone. Fig. 2 contains separate curves of total errors per attempted anticipation to the C(3) figure, on the one hand, and to all three C(1) figures, on the other. Errors and attempted anticipations were summed over all *S*s for each block. Zero errors were assumed for *S*s who had reached the criterion of two perfect blocks prior to a given block. It is apparent that most *S*s learned the name of C(3) within the first two blocks and spent the remainder of their learning time in discriminating among the C(1)s.

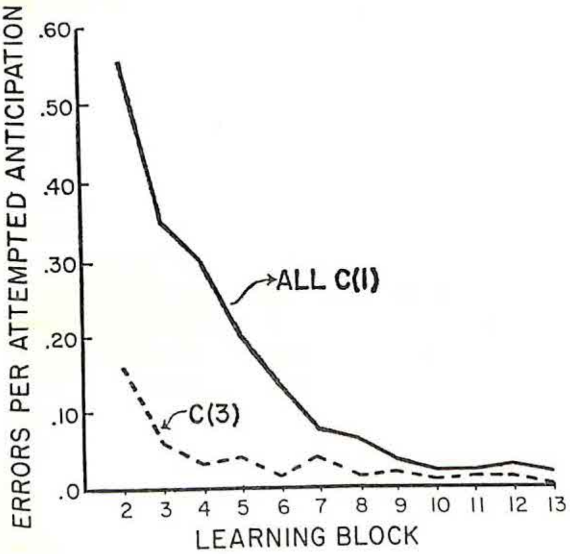


Fig. 2. Total errors per attempted anticipation over learning blocks for *S*s in Group II-C.

This would imply that for many *S*s there was more immediacy and fewer cue-producing responses involved in the association between the C(3) stimulus and its naming response, at least at certain stages of learning. Let us use an example to illustrate this argument. Suppose the paired associates for a given *S* in Group II-C are as follows: C_1—WOY, C_2—RUX, C_3—KAL, and C_4—GEP (see Fig. 1). Suppose, further, that C_3 occurs three times per block for this *S*, and C_1, C_2, and C_4 once each. Now this *S* quickly learns to respond "KAL" when C_3 is exposed. If this association is not immediate and such intermediate implicit responses as, "triangle, wiggly line, line to left" occur, they quickly drop out with the overlearning concomitant with the frequent repetition of only one stimulus. The essential body of learning consists of establishing discriminations among C_1, C_2, C_4. But this may require a good deal of emphasis and drilling on cue-producing responses. Thus, the learning may involve drilling on sequences like: C_1 exposed—"circle in middle and line in left corner"—"WOY"; or C_4 exposed—"triangle in center and bottom line to right"—"GEP." If this process of discriminating among the C(1) with the help of implicit responses proceeded through the learning trials for these *S*s, the difference in immediacy between responses to these stimuli and the response of "KAL" to C_3 would, up to a point, become more and more pronounced. When the Test Stimulus C is presented after that course of learning, a certain *S* might respond in some such fashion as: "No circle in the middle and no line in left corner, it is not a WOY; no line in right corner, it is not a RUX; no line in right corner and no triangle in center, not a GEP." The elements to which the cue-producing responses leading to the responses of "WOY," "RUX," and "GEP" are, in short, not present. The *S* then gives the response "KAL," the name of the stimulus which appeared three times per block.

With enough *S*s of Group II-C behaving in accord with the above model, the disproportionately large number of C(3) names becomes understandable. Unfortunately there was no way of evaluating the reasonableness of these formulations with the data thus far available and so new data had to be collected with proper experimental variation.

EXPERIMENT III

The purpose of this experiment was to test the hypothesis that the discrepancy between obtained and expected response frequencies for the *S*s in Group II-C resulted from the unusual distinctiveness of the particular figure occurring three times as often as the others per learning block. According to the analysis, this distinctiveness resulted in the *S*s learning the more frequent figure–name association comparatively directly and the less frequent figure–name associations with the aid of intervening implicit responses. And this in turn, presumably, led to a larger than expected frequency of the associated name of the C(3) figure on tests with Stimulus C.

The method used in this experiment involved a replication of the conditions used for the *S*s of Group II-C with one change introduced to reduce the distinctiveness of the figure which occurred three times per block. This diminution in distinctiveness was accomplished by increasing the frequency per block of two of the other three figures.

Method

Subjects. Sixty-five *S*s were used in this experiment. As in the previous experiments they were run individually and all were volunteers from introductory psychology sections.

Procedure. The apparatus and stimulus and response materials were the same as those described previously. Except for the specific composition of each learning trial block, the procedure used for the *S*s in Exp. III was identical to that used for the *S*s of Group II-C. Whereas each block for the *S*s of Group II-C consisted of one C_j presented three times and the other three once each, the blocks in the present experiment consisted of one C_j presented three times, another presented once, and the remaining two figures presented twice each. The learning blocks for the *S*s of Exp. III were, consequently, eight pairings in length. The particular figures among C_1, C_2, C_3, C_4 presented once, twice, and three times per block were chosen randomly for a given *S*, but once selected the particular figure frequencies were of course maintained through all of his learning blocks.

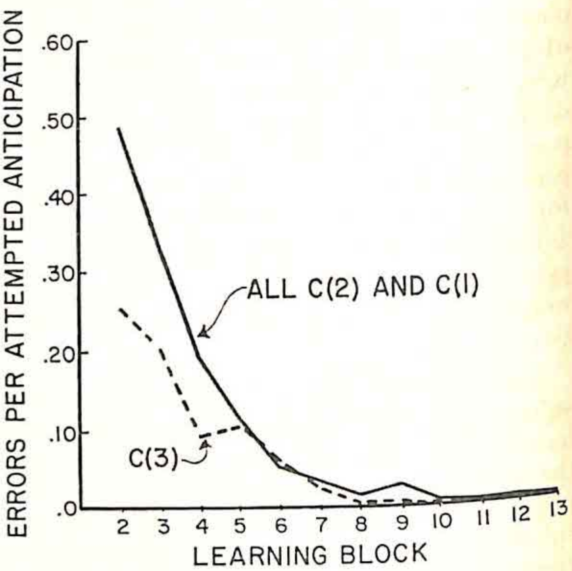


FIG. 3. Total errors per attempted anticipation over learning blocks for *S*s in Exp. III.

Results and Discussion

The condition whereby two learning figures occurred twice each per learning block surely would be expected to diminish the distinctiveness of the figure which occurred three times. That this occurred is indicated by the fact that 34% of the *S*s made no C(3) errors throughout their learning trial blocks; 37% made no C(3) errors on blocks two, three, or four; and 54% made none on block two alone. (The corresponding percentages for the *S*s in Group II-C are 49, 59, and 66.) Fig. 3 shows (for the *S*s in Exp. III) curves of errors per attempted anticipation to the C(3) figure, on the one hand, and the pooled C(2) and C(1) figures, on the other hand. Ordinate values were determined as in Fig. 2.

Let us evaluate the effects of this diminished distinctiveness. Table 8 contains the

TABLE 8

OBTAINED AND EXPECTED (IN PARENTHESES) RESPONSE FREQUENCIES TO TEST STIMULUS FOR *S*s IN EXP. III

C(3)	Both C(2)	C(1)
25	31	9
(24.4)	(32.5)	(8.1)

obtained and expected response frequencies of the *S*s in Exp. III to Test Stimulus C. Responses consisting of the name of either of the two C(2) figures were necessarily pooled since, considering that the figures presented at each frequency were random for a given *S*, there was no way of distinguishing between them in a frequency analysis over *S*s. The expected frequencies were computed by the procedure described above.

The closeness of the fit between expected values and obtained results is remarkable indeed. This supports the hypothesis that the discrepancy between obtained and expected values for the *S*s in Group II-C resulted from the unusual distinctiveness concomitant with the single figure's comparatively large number of occurrences per block. This does not, of course, provide powerful confirmation to the steps of the argument wherein the unusual distinctiveness is assumed to lead to the differential use of cue-producing responses and these in turn to a preponderance of test responses consisting of more frequent figure names. But there would seem to be no more compelling line of reasoning available to account for the discrepancy. Moreover, the presented line of reasoning was heuristic, and any other would be ex post facto.

The accuracy of prediction for these data make it seem desirable to determine if the obtained results are consistent with the independence of expected outcomes and the effects of recency. Table 9 contains this analysis. A given entry for predicted proportion is simply the product of the probability that the particular figure will occur last (or not last) during the final learning block and the probability, assuming the matching relationship holds, that the name of the figure will be given on the test trial. For example, the probability that C(3) is the last figure exposed is .375 and the probability that it will be named according to the matching hypothesis is .375; hence, the predicted proportion of C(3) responses when it is last is (.375)(.375) = .141. Again, the fit between predicted and obtained values is very close, indicating both the accuracy of prediction from the matching hypothesis as well as the relative unimportance of recency effects.

TABLE 9

OBTAINED AND EXPECTED PROPORTIONS OF RESPONSES IN RECENCY COMPARISON FOR EXP. III

	Test Stimuli					
	C(3)		Both C(2)		C(1)	
	Last	Not-Last	Last	Not-Last	Last	Not-Last
Obtained	.139	.246	.092	.385	.046	.092
Predicted	.141	.234	.125	.375	.016	.109

Thus, the results of Exp. II and Exp. III taken jointly provide very convincing support for the earlier finding of a direct relationship between a learning variable and a perceptual variable. Moreover, these results provide a relatively clearly defined set of conditions under which the relationship breaks down.

EXPERIMENT IV

The purpose of the present experiment was to provide one test of the hypothesis that, in a recognition situation, discriminative and common cues differ in the extent to which they are effective in determining responses because of differential sampling. Such differential sampling or, in other terms, selective attention has been hypothesized by certain theorists to account for perfect, or near perfect, discriminations in situations involving overlapping cues and common background stimulation.

In their application of statistical learning theory to the analysis of discrimination learning, for example, Estes and Burke (1955) found evidence for a reduction of common elements in the stimulus situation over the course of learning. La Berge and Smith (1957) explicitly pointed out that the assumptions of statistical learning theory imply that an individual making perfect discriminations in two-choice learning should be selectively ignoring common elements. They found support for the hypothesis that background common cues are ignored, in an experiment involving changes in partial reinforcement schedules with interspersed test trials consisting of only background elements.

Wyckoff (1952) used the term "observing responses" to designate those responses of an *S* which are necessary for exposure to the discriminative stimuli. Wyckoff presented an analysis of changes in the probabilities of these responses in terms of reinforcement learning theory, equating these changes with alterations in stimulus generalization between the stimuli. Atkinson (1959) dealt with observing responses within the framework of a Markov model, in an approach related to the earlier work of Wyckoff. He modified the typical discrimination task significantly by equating observing responses with the operation of selection keys, in order to make the process of selective attention directly measurable. Restle (1955), on the other hand, used Lawrence's (1950) concept of "irrelevant cues," which are those aspects of a stimulating situation which are unrelated to reinforcement. Restle hypothesized that these irrelevant cues become progressively adapted over the course of learning, and thus eventually cease to influence the discrimination responses of *S*s. Recently, Restle (1957) has presented a more general theory of discrimination learning in which rate of adaptation is a function of the information-giving value ("validity") of the available cues. Finally, Green (1958) described the process of selective sampling of stimulus elements in terms of a reduction of the intercept or common set as a result of postural or orientation adjustments.

In summary, these investigators have attempted to account, by means of such concepts as "observing responses" and "cue adaptation," for the process whereby *S*s learn to ignore the common (or nondiscriminative or irrelevant) cues in discriminating among stimuli. The term "selective sampling" will be used to designate this process throughout the present discussion since the expressions "observing responses," "cue adaptation," and "cue neutralization" carry connotations which are not intended. "Selective sampling" will imply no more than a designation of conditions, at a given discrimination point, whereby certain cues are effective in influencing responses while others are not.

In order to test the hypothesis that selective sampling occurs in recognition situations of the sort investigated in the preceding experiments of this project, use was made of the finding that the relative frequencies of responses to ambiguous cues tend to equal the relative frequencies of occurrence of these responses in previous S–R associations. Current research would indicate that selective sampling occurred in the preceding experiments, since the majority of *S*s learned to discriminate among the figures during the learning trials with no errors after the initial criterion of two perfect blocks was reached. The postulate in statistical learning theory (Estes & Burke, 1955), for example, that the probability of a response on a given trial is equal to the proportion of conditioned elements leads to the assumption that the common elements in both the figures and the background were not sampled (that is, were ignored) in the process of response determination.

Paired-associates learning was again used in this experiment in the manner of the previous experiments. The present experiment differed from the others, however, in that the relative frequencies of presentation of the learning stimuli per trial block were changed once during the course of learning. The purpose of the shift was to introduce a different rate of conditioning of the stimulus elements. For those *S*s for whom selective sampling had been operative at the time of the shift, the ignored cues would not be conditioned at the new rate. A determination of whether common cues were responded to at the initial relative frequency or at the relative frequency following the shift then provided a means of assessing if selective sampling had actually occurred.

Method

Apparatus. The exposure of stimuli and responses was accomplished by the apparatus described in Exp. I.

Subjects. Two groups of 38 *S*s each were used in the experiment. They were volunteers from the introductory psychology sections at Indiana University.

Materials and procedure. Four learning figures and three test stimuli were used; these may be seen in Fig. 4. A paired-associates procedure was used during the learning phase, in which the figures shown in the left hand column of Fig. 4 were associated with the nonsense syllables RUX, TUZ, WOY, JOM. The matching of figures and syllables was done randomly over *S*s. The sound of a buzzer preceded each stimulus presentation.

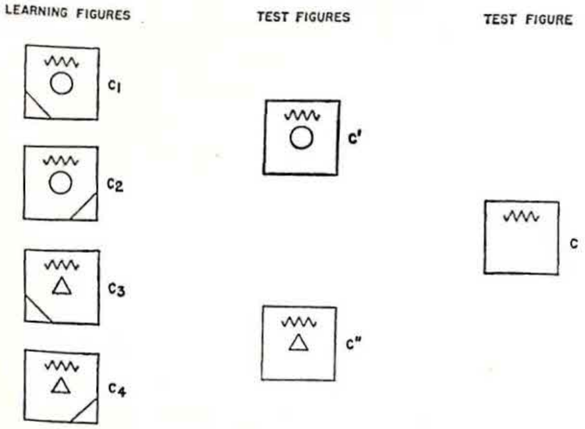


Fig. 4. Learning and test figures for Exp. IV.

The timing of presentations and intertrial intervals was the same as that of Exp. I.

For the first three trial blocks, each stimulus response pair was shown once per block. Then, assuming that a set of three learning blocks is a reasonable interval in which to expect selective sampling to begin occurring, a differential frequency of presentation per block was used beginning with the fourth block and continuing through the learning phase. During these latter blocks, C_3 or C_4 occurred twice and the other once, while C_1 or C_2 appeared four times and the other once. While C_3 and C_4 always appeared in the ratio of two to one per trial block, the particular one which occurred twice and the one which occurred once were randomly determined. In a similar fashion, whether C_1 or C_2 appeared four times per block was randomly determined for different *S*s. The order of presentation was random on odd blocks and the reverse of the previous odd block on even blocks.

Upon achieving two trial blocks with perfect anticipations, the *S*s were given a 3-min. rest period. Then, during the test phase of the experiment, they were required to achieve one further perfect trial block before each test trial. The test trials consisted of the presentation of C, C′, and C″ (see Fig. 4) to all *S*s. Test Stimulus C′ contains a cue set common only to Figures C_1 and C_2, Stimulus C″ a cue set common only to C_3 and C_4, but Stimulus C contains nothing but cues common to all four learning figures.

The two experimental groups differed in the ordering of the test trials. The Group IV-C′ *S*s were tested on either Stimulus C′ or Stimulus C″ first, the remaining one of these two stimuli second, and Test Stimulus C last. The Group IV-C *S*s were shown C first, and then C′ and C″ randomly ordered, second and third.

For the learning instructions, the *S*s in both groups were told that the nonsense syllables were to be considered the names of the figures and that the syllables were to be pronounced in the process of anticipation. Prior to each test trial all *S*s were told that a new figure would be presented and that their task was to respond with the one previously learned name which seemed most appropriate.

It became apparent after these *S*s were run that a control group was needed. This was needed because no previous data were available for *S*s tested on a stimulus like C after a 2:1:4:1 frequency throughout learning (that is, starting with 2:1:4:1 during the first block and maintaining that frequency on all subsequent blocks). Accordingly, a group of 29 *S*s was run with C_1 and C_2 shown in a 4:1 ratio and C_3 and C_4 in a 2:1 ratio throughout learning (see Fig. 4). The test stimulus consisted of C alone.

Results and Discussion

If an *S* gives an appropriate response to Test Stimulus C′ (that is, the name of C_1 or C_2), his subsequent response to C″ must be different (that is, the name of C_3 or C_4) or else it will be inappropriate. And the same holds, of course, for the sequence C″ and then C′. In that sense, the responses to C′ and C″ are necessarily independent for any analysis involving specific appropriate responses. In contrast, the same responses are appropriate to either C′ or C″, on the one hand, and C, on the other. Thus, it is possible to have sequential dependencies in terms of recurrences of the same response which is appropriate for both stimuli. The previous experiments avoided the problem of possible lack of independence by using only those test stimuli (like C′ and C″) in the group administered to a given *S*, for which the associated sets of appropriate responses were nonoverlapping.

The present investigation was so designed that the effects of using successive test stimuli for which the same responses were appropriate could be determined with little loss of experimental time. All *S*s were shown the three stimuli C, C′, and C″ in such a fashion that C came either first or last. Thus, if it were found that the response patterning to the C′ and C″, on the one hand, and C, on the other, are independent, the combined data could be used; but if no such independence were shown, the data for the *S*s seeing C first and those seeing it last could be analyzed separately, which would have been necessary even if all test stimuli had not been shown to each *S* as in the present investigation of independence.

Table 10 contains the response frequencies to the test stimuli for the Group IV-C′ and Group IV-C *S*s. The response patterning would seem to be quite different for the two groups, with a distinct tendency for lower frequency responses to occur to C when C′ or C″ came first and to C′ and C″ when C came first. Thus, the ratio of more frequent to less frequent responses is 2:1 for C″ and 2.5:1 for C′ in the case of the Group IV-C′ *S*s, and the corresponding ratios are 1.5:1 and 1.7:1 for the Group IV-C *S*s. These data may be compared with those for Group II-AB because of the similarity of learning frequencies and test conditions in the two situations. The ratio of more frequent to less frequent responses for the *S*s in Group II-AB is 2.2:1 to Test Stimulus A (comparable to C″) and 3.8:1 to Test Stimulus B (comparable to C′). This comparison adds some weight to the evidence of nonindependence in test responses when the same response is appropriate to successive test stimuli, despite the marked discrepancy in response tendencies between C′ and B. It is surely clear that the response ratios to C′ and C″ are much closer to the results for the *S*s in Group II-AB when C′ and C″ appear prior to C than when they appear after C.

In the case of the responses to Test Stimulus C, there is an increase, when C comes last, in the responses which occur at frequencies of two and one at the expense of those of frequency four (see Table 10). The results from the group of 29 control *S*s are relevant here. These were: 16 responses associated previously with the figure occurring four times, 7 responses which occurred twice, and a total of 6 of the two responses which occurred once each. These data are much more nearly like the results when C was presented first than when it was presented last.

All the data from this experiment, as compared with the results for Group II-AB and for the control group of 29 *S*s, indicate that the learning blocks of Exp. IV subsequent to the third were not completely effective in overcoming the response tendencies stemming from the three initial blocks of equal frequency. That is, when the four learning figures are presented equally frequently for three blocks and then shifted to a 2:1:4:1 frequency, there is an increase in the relative frequencies during testing of naming responses of figures which occurred only once per learning block, as compared with the results when no shift occurs and the frequency 2:1:4:1 is main-

TABLE 10

RESPONSE FREQUENCIES (FOR EXP. IV) IN TERMS OF THE LEARNING FREQUENCIES OF ASSOCIATED STIMULI

				C″			C′			C	
Group	Number of *S*s	First Test Stimulus	2	1	Error	4	1	Error	4	1	2
IV–C′	38	C′ or C″	22	11	5	27	11	0	9	17	12
IV–C	38	C	19	13	6	22	13	3	17	13	8

Note.—Numbers in columns headed by C″, C′, and C refer to frequency of occurrence, during learning, of named figure, if appropriate.

tained throughout learning. Moreover, as pointed out above, there is a further increase in the frequency of test responses consisting of the names of figures which occurred once per block when the test stimulus followed (for a given *S*) other test stimuli having the same appropriate responses. This latter increase may have resulted from the attempts of *S*s to use as many different testing responses as possible. That is, there were obviously more high than low frequency responses initially and so any general tendency toward using different responses on subsequent test trials would lead to a net shift in the direction of low frequency responses.

In any event, as a result of the indicated lack of independence, only the *C′* and *C″* test responses of Group IV-C′ and the C responses of Group IV-C were used in the analysis of selective sampling.

Table 11 contains the distribution of errors made subsequent to the first required criterion level of two blocks with perfect anticipations. It should be remembered that all *S*s were required to achieve one further perfect block before each test trial. The deviation from perfect responding during criterion trials may conceivably have resulted from inadequate selective sampling of discriminative cues. Consequently, each of the two groups was divided into two parts: one part contained those *S*s who made no errors whatsoever in the learning blocks following the first two required perfect ones, and the other part contained *S*s who made one or more errors during these blocks. For ease of designation, Group IV-C′-NE shall refer to those *S*s who were shown Stimulus C last and made no errors during the criterion blocks, Group IV-C′-E, to those *S*s who were shown Stimulus C last and made one or more errors during the criterion blocks; and Group IV-C-NE and Group IV-C-E shall have similar meanings.

TABLE 11

Total Errors of Anticipation Made by *S*s in Exp. IV in Learning Blocks Following First Perfect Consecutive Pair

Errors	Number of *S*s
0	44
1	15
2	12
3	2
4	1
5	1
6	1

The curves, over all learning blocks, of average number of errors per *S* in each of the four subgroups may be seen in Fig. 5. There is some difference in learning trend evident between the Group IV-C′-NE and Group IV-C′-E *S*s, on the one hand, and the Group IV-C-NE and Group IV-C-E *S*s, on the other. It is hardly surprising that the "NE" *S*s as a group were more efficient learners than the "E" *S*s. There would seem to be little difference between the learning rates of the *S*s in Group IV-C′-NE and those in Group IV-C-NE during the later blocks, and these will be the critical comparison groups in the analyses of selective sampling. The average number of blocks to reach the final necessary perfect criterion block for each of the subgroups is as follows: Group IV-C′-NE, 11.24; Group IV-C′-E, 12.61; Group IV-C-NE, 10.58; Group IV-C-E, 12.95. The difference between the Group IV-C′-NE and Group IV-C-NE *S*s is small and far from significant.

We would expect, if selective sampling occurs, that it does so for the *S*s at high

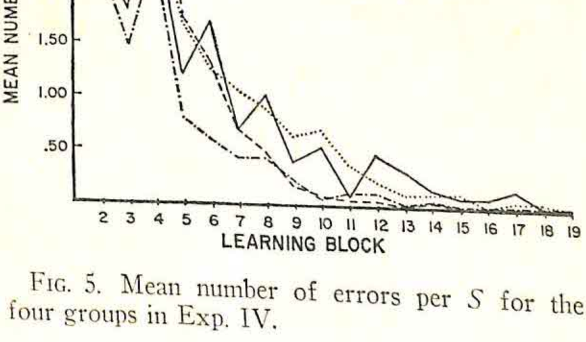


Fig. 5. Mean number of errors per *S* for the four groups in Exp. IV.

learning levels who show perfect discriminations over the criterion blocks.

Table 12 contains a summary of the test responses of the various subgroups in terms of the frequencies of occurrence of the associated learning figures. Responses consisting of the name of a figure which occurred twice or four times per block have been pooled under the "More Frequent" heading, while those consisting of a name occurring once per block have been pooled under "Less Frequent." All erroneous test responses were eliminated from consideration. The pooling into these two categories was done because of the relatively small frequencies which would have resulted from the finer breakdown involving the actual numbers of learning occurrences.

Expected frequencies for the pooled data can be computed by the usual rules for combining fractions. Thus, for example, the "More Frequent" responses should occur three times as often as the "Less Frequent" responses if the *S*s in Group IV-C-NE responded to the test stimuli at the same relative frequencies as the occurrence of the relevant figure–name pairs during later learning. If selective sampling had occurred prior to the fourth block, however, an equal number of C responses in the "More Frequent" and "Less Frequent" categories would be predicted. That is, if the common cues were not sampled as a result of selective attention, they could not be conditioned to the various responses at the differing frequencies.

Stimuli C′ and C″ contain discriminative cues which could not be ignored if *S*s were to give correct anticipations during learning. Hence, approximately a 2.82 to 1 ratio of "More Frequent" to "Less Frequent" responses was expected for the *S*s in Group IV-C′; the expectation is approximate rather than exact because of the previously noted ineffectiveness of the shift to 2:1:4:1 in completely overcoming the response tendencies stemming from the first three blocks of equal freqency. The test for the hypothesis of selective sampling involved the Group IV-C-NE *S*s, since they responded to the test stimulus containing only common cues and showed perfect discriminations at criterion trials. Table 13 shows observed and expected values for the *S*s in Group IV-C′-NE and also for the *S*s in Group IV-C-NE.

The observed and expected values do not differ significantly for the Group IV-C′-NE *S*s ($P > .25$) using a χ^2 test. The observed values for the Group IV-C-NE *S*s differ significantly from that expected if they did not selectively sample ($P \leqq .05$), but do not differ significantly from the values expected under the assumption of selective sampling by the fourth block ($P = 1.00$).

It is interesting to note further (see Tables 12 and 13) that the *S*s in Group IV-C-E responded very close to the 3:1 learning ratio (14.25 to 4.75 expected vs. 15 to 4 observed). It is surely clear that the Group IV-C-E *S*s did not show selective attention. It is conceivable that their comparative learning inefficiency resulted from their deficiency of selective sampling among common and discriminative cues. Noting the response ratios of the Group

TABLE 12

RESPONSE FREQUENCIES IN TERMS OF LEARNING FREQUENCIES OF ASSOCIATED STIMULI FOR EXP. IV

Group	Performance After First Criterion Blocks	Number of *S*s	Test Stimuli	Learning Occurrence of Stimulus–Response Pair: More Frequent	Learning Occurrence of Stimulus–Response Pair: Less Frequent
IV–C′–NE	No errors	25	C′ and C″	33	14
IV–C′–E	One or more errors	13	C′ and C″	16	8
IV–C–NE	No errors	19	C	10	9
IV–C–E	One or more errors	19	C	15	4

TABLE 13

OBSERVED AND EXPECTED RESPONSE FREQUENCIES FOR *S*s OF EXP. IV

			Learning Occurrence of Stimulus–Response Pair	
Group	Value Source		More Frequent	Less Frequent
IV–C′–NE	Observed		33	14
	Expected	(Differential Frequency Matching)	34.67	12.33
IV–C–NE	Observed		10	9
	Expected	(Selective Sampling)	9.5	9.5
	Expected	(Differential Frequency Matching)	14.25	4.75

IV-C-E and the Group IV-C′-NE *S*s in Table 12 (3.75 : 1 and 2.36 : 1, respectively), the relatively low number of "More Frequent" responses on the part of the IV-C′-E *S*s becomes difficult to account for on any basis other than sampling fluctuation. If the data for the *S*s in Groups IV-C′-E and IV-C-E are combined, they give a total of 31 "More Frequent" and 12 "Less Frequent" responses. This is very much like the response ratio for the Group IV-C′-NE *S*s.

To summarize, this experiment provided a test of the hypothesis that responses to ambiguous cues are a function of differential attention to discriminative as compared to common stimulus elements during the process of learning the names of relevant stimulus figures. The test was based on the usual presumption that learning only occurs with cues actually sampled. This being the case, it was assumed possible to detect cessation in the sampling of certain cues by an initial learning experience followed by an altered learning experience subsequent to the assumed cessation. If selective sampling occurred to the given cues, test responses to them at the end of the learning period should have been in accord with the initial learning; while, if selective sampling had not occurred, the responses should have been in accord with altered learning.

Only *S*s at high criterion levels of learning were used since it appeared likely that lower levels of learning may have resulted from poor selective attention. It was found that *S*s who were shown discriminative cues responded in accord with the later (altered) learning experience, while *S*s shown only common cues responded in accord with the initial learning experience. This supported the hypothesis of selective sampling in perceptual-type situations involving common and discriminative cues.

The data are in accord, moreover, with the findings of La Berge and Smith (1957) that a change in the partial reinforcement schedule associated with the common cues shown on test trials produces no change in the level of response for *S*s near asymptote but does produce a change for *S*s further from asymptote. While the theory of Restle (1955, 1957) on the adaptation of common cues accounts for the general findings, it does not contain a sufficiently clear formulation of the relationship of adaptation to conditioning. Thus, the *S*s in the present experiment responded quite accurately to the common cues which had presumably been previously adapted. Restle, to be sure, recognizes the phenomenon in his recent statement (1959) that "irrelevant cues are 'neutralized' relative to the relevant cues present, and recover if the relevant cues are removed or scrambled" (p. 14) as well as in his previous comment (1958) that, when certain cues are removed, invalid cues recover and performance tends to change. But it is not apparent just where these observations fit into his more formal theory of discriminative learning.

It would seem more economical at the present time to consider the sampling and conditioning processes separately in accord with the formulations of Estes and Burke (1953). In this way a given stimulus element may be conditioned to a response but be ineffective in eliciting the response because it is not sampled. The probability of sampling the given element may be thought of as going to zero during the process of selective attention. When the stimulus situation is changed sufficiently so that the element is again sampled, its previous conditioning is still effective.

EXPERIMENT V

In all of the experiments reported so far the *S*s learned to associate nonsense names with nonsense figures which contained overlapping cues. Following the learning, the *S*s were required to name a test stimulus which contained only cues which were common to two or more of the previous learning figures. There was, in the terms used in the introduction, ambiguity as a result of reduction of information when the full set of cues was necessary for unique discrimination. Moreover, the test situation and analysis were devoid of phenomenological or "autistic" considerations. Surely it would not be easy to argue that an *S* who called a given test stimulus a RUX "really saw" the figure with which that name had been associated. Such phenomenological constructs could be anchored by other aspects of the behavior of the *S* as well as by appropriate questioning. These aspects of perception will be discussed more fully later.

Nevertheless, these considerations led to Exp. V. The test stimuli of Exp. V differ from those of Figs. 1 and 4 in that their ambiguity results from too many, rather than too few, cues; and in addition, there is probably more of a "really see" reaction associated with responses to them. Perhaps these should not be listed as two factors because of their close interdependence, but it is convenient to do so to highlight the separate aspects.

The learning and test stimuli used in this experiment were based on those employed by Schafer and Murphy (1943). These experimenters hoped to demonstrate that perception was influenced by "needs." They attempted to show that, when *S*s were required to name an ambiguous presentation consisting of a combination of a previously rewarded and previously punished profile, the latter would form the ground while the former would stand out as figure. The stimuli used by Schafer and Murphy were identical to those in Fig. 6 except that each of their learning figures also included the eye dot of the profile that would be its compliment when it was presented as part of a test figure.

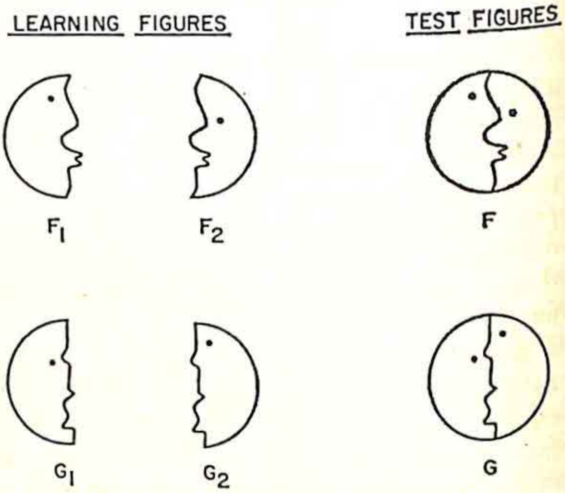


FIG. 6. Learning and test figures for Exp. V.

In the Schafer and Murphy experiment *S*s were shown the four half-moon faces and were required to learn their names. For each *S* one face in each pair (such as F_1 or F_2) was regularly rewarded (2 or 4 cents given to *S*) while the other face was regularly punished (2 or 4 cents taken from *S*). The test stimuli of Fig. 6 which contain the combination of both figures of a pair, one rewarded and one punished, were then shown to the *S*s under instructions to name them. Fifty-four rewarded as opposed to 13 punished names were given to the test stimuli.

Rock and Fleck (1950) repeated the Schafer and Murphy study but failed to confirm the previous findings. However, as pointed out by Jackson (1954), there were several changes in the Rock and Fleck procedure that may have accounted for the differences in results. Among the changes were: younger *S*s, the use of symbols rather than proper names, the change in "reward value" of 2 and 4 cents since 1943, and the use of a projection tachistoscope which resulted in larger faces, lessened clarity, and the absence of a constantly lighted field. Jackson repeated the Schafer and Murphy study with some modifications, including increased rewards, and once again showed that rewards and punishments had their predicted effects. However, when he substituted a projector, for presentation purposes, with a second group of *S*s, his findings concurred with those of Rock and Fleck.

Smith and Hochberg (1954), substituting 15-in. solid figures for the outline profiles, demonstrated that a shock vs. no shock condition resulted in findings similar to those obtained for monetary reward and punishment. *S*s gave a significantly larger number of nonshocked figure names when they were required to name ambiguous test figures.

More recently, Solley and Sommer (1957) replicated the Schafer and Murphy experiment using children, increased monetary rewards, and solid profiles. They obtained a significantly greater number of rewarded names when their *S*s were shown the ambiguous test presentations for one-min. periods and were asked to describe what they saw. In a second experiment, these authors found that *S*s reported the rewarded profiles as "happier," "brighter," "closer," and as having a dark line drawn around them. Solley and Sommer concluded that reward "produced a genuine figure-ground reorganization" (p. 9).

Considering these results the purpose of this experiment may now be stated: (*a*) to test predictions of the relationship between relative frequency of S–R association and relative frequency of naming identification and (*b*) to compare the effects of frequency upon responses with those of rewards and punishments. In addition, because of certain modifications introduced in the present experiment, some information about the generality of previous findings might be established.

Method

Subjects. The *S*s were 66 volunteers from the courses in introductory psychology at Indiana University. The learning criteria were not met by two of these *S*s and, therefore, the data from 64 *S*s were included in the analyses.

Apparatus. The apparatus employed was essentially the same as that described in Exp. I. The only important difference was in the size of the exposure mirrors, which were 3 in. by 3 in. in the present experiment.

Learning and test materials. Each *S* was required to learn the nonsense names of four half-moon faces. The four faces used for any *S* were either direct photographic reproductions of those faces employed by Schafer and Murphy with the eye dot exception as noted previously (see Fig. 6) or photographic reproductions of the negative images of these faces. This pool of eight figures was such that figures could be paired on the basis of shared common contours. The test figures consisted of these paired figures. Thus, if F_1, F_2, G_1, and G_2 were shown during the learning phase, the pairings F_1 and F_2 to produce F and G_1 and G_2 to produce G were the two test figures.

The test figures were 2½-in. photographic reproductions mounted on cardboard which, upon presentation, were seen centered in the appropriate window. A given face, whether it appeared individually or as part of a test figure, never varied from a set location of exposure. The nonsense syllables were printed in ½-in. india ink capitals, on cardboard, such that on presentation they were seen as centered in their window. The four nonsense syllables used were WOY, JOM, RUX, and TUZ.

Procedure. The procedure was similar to the one used in the previous experiments. However, the following intervals were used: exposure of stimulus 4 sec., exposure of response 2 secs., interval between trials 6 secs., and delay after warning buzzer 1 sec.

Eight trials constituted a learning block. On any block, one member of a pair containing common contours appeared four times while the other member appeared once. The ratio of the frequencies of presentation of the members of the other pair was two to one. An example of the relative frequencies of presentation of the four faces for a given *S* may be seen in Table 14.

TABLE 14

NUMBER OF OCCURRENCES OF LEARNING FIGURES PER TRIAL BLOCK FOR ONE POSSIBLE ASSIGNMENT IN EXP. V

Learning Figure	Frequency
F_1	4
F_2	1
G_1	2
G_2	1

The order of presentation of the trials for each block was random. An attempt was made to control for possible biases for particular faces and directions by the use of a counterbalanced design, such that for a given *S* the particular group of pairs used, the pair that was shown in the four to one ratio, and the member of higher frequency within each pair were randomly determined.

The learning procedure continued until an *S* anticipated perfectly on two consecutive trial blocks. After a rest period of 3 min., the *S* was told that following the next perfect block he would be shown a new figure and would be required to give one of the names of the previously learned figures. Following this perfect block the *S* was reminded of his task and then shown a test stimulus. The test stimuli were shown for a duration of 1½ secs., beginning 1 sec. after the sounding of the warning buzzer. After the first test response, the *S* was required to anticipate perfectly another block and was then shown the second test figure. The order of presentation of the test figures was random, with the restriction that the figures corresponding to the 4:1 and 2:1 pairs were shown first an equal number of times.

Results and Discussion

Table 15 shows the observed and predicted frequencies of responses to the test stimuli. The column headed by "Four" contains the number of obtained and predicted responses designating the name of a face that appeared four times per learning block. The other columns correspondingly contain the number of responses designating figures which were shown at the frequencies shown in the column heading. The errors are inappropriate responses. Referring to Fig. 6, if the Test Figure G were shown, a response consisting of the name of F_1 or F_2 would be inappropriate. If the components of the Test Figure G had been presented in a four to one ratio, an error made to this figure would be represented in the third column, and if the ratio were two to one, in the last column.

The predicted frequencies refer to the number of appropriate responses expected on the basis of the hypothesis that the relative frequencies of appropriate responses to ambiguous cues are equal to the relative frequencies of presentation of the names associated with these cues during the learning phase.

The differences between expected and observed values (at least for the 4:1 comparison) are quite large, especially as compared with the very close fits obtained in the previous experiments. Using a χ^2 test, the observed and predicted frequencies differ at the .01 level in the 4:1 comparison and at the .20 level in the 2:1 comparison. The hypothesis of the matching of relative frequencies of responses and relative frequencies of occurrence during learning trials in this type of situation is certainly not well-supported by these data.

There are a number of differences between the present experimental conditions and the earlier ones in this project. First, the specific learning figures differ; second, the present figures are more meaningful than those used previously; third, the ambiguity in the test stimuli resulted from an increase of cues in the present case, while it resulted from an impoverishment of cues in the previous experiments; and fourth, while certain cues in the previous experiments were on a partial reinforcement schedule, all cues in the present experiment received 100% reinforcement. There is one other difference which, though less obvious than any of the preceding, seems of utmost importance. The configurations of cues in

TABLE 15

Observed and Expected Responses to the Test Stimuli in Exp. V

Learning Figures	Appropriate Responses		Inappropriate Responses	Appropriate Responses		Inappropriate Responses
	Four	One		Two	One	
Observed	40	21	3	32	23	9
Predicted	48.8	12.2		36.7	18.3	

TABLE 16

NUMBER OF *S*s[a] SWITCHING AND REPEATING RESPONSES ON THE TWO TESTINGS IN EXP. V

		Second Response Frequency					Second Response Frequency		
		Four	One	Row Sum			Two	One	Row Sum
First Response Frequency	Two	9	8	17	First Response Frequency	Four	13	8	21
	One	6	2	8		One	2	5	7
	Column Sum	15	10			Column Sum	15	13	

[a] Only *S*s who gave two appropriate responses are included in analysis.

the stimuli of the present experiment were such that the second test stimulus was obvious to an *S* once the first test stimulus had been seen. This was not the case in the previous experiments where the learning figures were nonsensical and the test configurations were arbitrary choices among the cues of these figures. Thus, while the first and second responses were independent in the previous work, the ability of an *S* to predict that the second compound face would be shown once the first had been exposed could lead to a lack of such independence.

Table 16 divides the appropriate test responses into categories defined on the basis of first and second responses. Thus, for example, it contains the number of *S*s who gave the names of the learning figures of higher frequency on both trials, among those who were presented on the first test trial with the stimulus whose components occurred in a 4:1 ratio and on the second trial with the stimulus containing 2:1 components (that is, 13). The results for the other combinations are similarly included. The differences between response tendencies on the first and second trials are not statistically significant (using Fisher's exact test). However, the very small sample sizes make it extremely unlikely that a significant difference would be found even if the obtained differences were much more extreme. Nevertheless, there are certain tendencies evident in the table: (*a*) a surprisingly large number of marginal responses of low frequency on the second test trial, (*b*) *S*s who give responses of low frequency on the first test trial to the compound face with components shown 4:1 give more responses of low than of high frequency on the second test trial, (*c*) the relative frequencies on the second test trials tend to vary with the particular first test response given (though not significantly).

A conclusion of nonindependence is consistent with the findings of Schafer and Murphy (1943). They encountered, in a pre-experimental phase, a tendency for *S*s to respond, on repeated testings, with the names of faces pointing in a particular direction. They attempted to counteract this effect by inserting, between each test trial, a configuration with only one acceptable face which pointed in the direction opposite that of the face selected on the previous ambiguous trial. However, they found, despite this attempt at obtaining trial-to-trial independence, that after approximately eight exposures of an ambiguous figure, equal percentages of rewarded and punished faces were given. They accounted for this finding in terms of a "new set" and consequently only considered the responses to the first half of their test presentations.

Pastore (1949) criticized Schafer and Murphy's analytical procedure since he felt that the cut-off point was a *post hoc* choice dictated by the fact that the hypothesis was supported only up to the selected point.

Chein, Lane, Murphy, Proshansky, and Schafer (1951) countered by saying that the choice of the sixteenth trial allowed a wide "margin of safety." They pointed out, further, that the responses after the sixteenth trial were not random, rather "there was clear evidence that *another process* intruded and took over" (p. 134).

Rock and Fleck (1950) referred to "a certain consistency which develops in responding to a situation where there is no check on correctness or incorrectness after each trial" (p. 775). They suggested that "a better, though practically infeasible, procedure would be to give, say, 100 *S*s each one trial with each ambiguous situation in order to avoid the contaminating effect of earlier presentations of ambiguous figures upon subsequent ones" (p. 775).

On the basis of these observations, the first test response of each *S* was considered separately for further analysis. Table 17 contains the observed and predicted responses to the first test stimuli. The deviations of observed results from those predicted are indeed small and very far from being statistically significant. Thus, there is some support for the hypothesis that even in a stimulus situation of the sort used in the present experiment, the relative frequencies of the various first test responses to the ambiguous cues tend to be equal to the relative frequencies of occurrence of these responses to their associated cues during learning. This does not hold for the second test responses, apparently, because of a lack of independence. Since the original predictions were made for both test trials and the reasoning toward lack of independence made a posteriori, these results must be cross-validated.

Considering the two testings separately, there are two types of biases that may have been operating to influence the responses given to the ambiguous figures. The first of these is a preference for giving the names of figures pointing in a particular direction. Jackson (1954) found a left-facing preference for the G figure and substituted a new profile set in his replication of the Schafer and Murphy (1943) experiment. Subsequently, he found no directional preferences. However, when he employed the Rock and Fleck (1950) projection procedure he found a left-facing preference. In the present experiment an analysis of directional preferences revealed a left-facing tendency for Test Figure F and its mirror image F′, and also for Test Figure G and its mirror image G′, although the preference was significant only in the former case ($P \leq .05$).

The second bias that might have been operating is a structural one: the naming of that face of the test figure that is more easily seen as a face, that is, "stands out more." For example, if F_1 were more readily seen as a face than F_2, its name would be given more often independently of learning presentation frequency and directional considerations (right facing in F, left facing in F′). In the present experiment no such preferences appeared, as none of the test figure presentations produced a significantly greater proportion of names of a particular profile than could be expected by chance ($P > .30$, $P > .20$ for F and F′, and G and G′, respectively).

In summary of the work with figures of the Schafer-Murphy type, a number of investigations have demonstrated that per-

TABLE 17

OBSERVED AND EXPECTED RESPONSES TO FIRST TEST STIMULI IN EXP. V

Learning Figures	Appropriate Responses		Inappropriate Responses	Appropriate Responses		Inappropriate Responses
	Four	One		Two	One	
Observed	22	7	3	17	8	7
Predicted	23.2	5.8		16.7	8.3	

ceptual responses could be predicted by an analysis of the previous learning situation in terms of rewards and punishments. The present analysis, however, not only enables us to predict a significant result in a given direction but also enables us to predict an exact quantitative relationship. And while this prediction held for the first testing responses only in the present case, the procedure of analysis involving the discarding of nonindependent data is the same as that used by Schafer and Murphy (1943) (who also discarded half their test); in fact, this mode of analysis conforms to the suggestions of Rock and Fleck (1950) who believed this scheme would provide the most cogent test of the hypothesis under consideration.

The evidence from these experiments, then, points toward the importance of frequency of S–R pairing during learning in determining responses to ambiguous cues. In fact, it indicates, under quite general conditions, a precise quantitative relationship between the variables learning frequency of figures and response frequency to figure parts. But, as discussed above in the introduction, there are many other factors which apparently influence the choice of a particular response out of a larger group of equally accurate responses to a set of ambiguous cues. Some of these are food deprivation (Sanford, 1936; Sanford, 1937; Levine et al., 1942), monetary reward and punishment (Schafer & Murphy, 1943; Proshansky & Murphy, 1942; Solley & Sommer, 1957), electric shock (Smith & Hochberg, 1954), and pleasant and unpleasant stories (Solley & Sommer, 1957). The remaining experiments in this project were directed toward investigation of a few of these types of differential reinforcement.

EXPERIMENT VI

Recent studies in operant verbal conditioning have shown that the response of "Good" on the part of the experimenter is effective in producing a significant increase in selected response categories. Thus, for example, Taffel (1955) found that the response "Good" was an effective reinforcer in leading to a significant increase in the use of the particular pronouns with which it was associated. And Binder, McConnell, and Sjoholm (1957) found that "Good" was reinforcing in that it led to an increasing usage of "hostile" rather than "neutral" verbs in sentences. (For a thorough review of the literature on verbal conditioning see Krasner, 1958.) The conditioning effect in these studies was found even though the tasks were not presented as learning ones and the *S*s were unaware of both the reinforcement contingency and the actual learning process. In the present series of experiments, interest was of course not centered in acquisition during the learning phase but rather in the choice among responses to reduced cues subsequent to the achievement of criterion. The question was whether or not the experimenter's response of "Good" would produce a test bias in favor of the figure–name pair with which it was previously associated, much as it led to a bias toward a particular class of responses during the acquisition in verbal conditioning.

Method

Subjects. Twenty-nine *S*s from the introductory psychology sections at Indiana University were used.

Apparatus and materials. The same apparatus as that discussed in Exps. I, II, and III was used. The stimulus materials may be seen in Fig. 7. The dimensional and other specifications for the stimuli may be found in the *Methods* section of Exp. I. The response syllables associated with the stimuli were JOM, VED, NAF, and RUX.

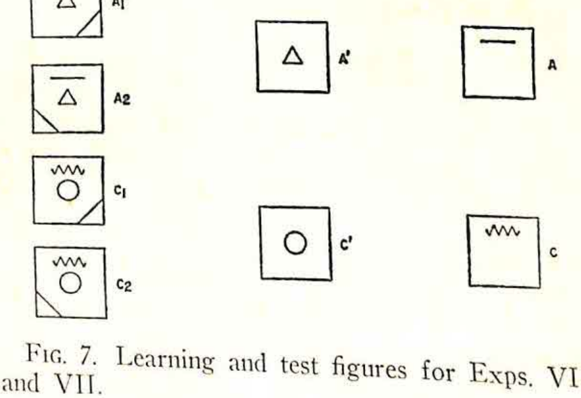


Fig. 7. Learning and test figures for Exps. VI and VII.

Procedure. The figures and their associated nonsense syllables were shown to the *S*s by the method of paired associates. Figures and names were paired randomly over *S*s. Except for the experimenter saying "Good" after certain of the *S*s' responses, the learning trials proceeded in the manner of the previous experiments.

A learning block consisted of a single appearance of each stimulus and its response. The experimenter said "Good" immediately after *S* gave the correct response to either A_1 or A_2, on the one hand, and either C_1 or C_2 on the other. While the particular stimulus in each pair (that is, A_1 or A_2 and C_1 or C_2) associated with "Good" varied randomly over *S*s, all *S*s were reinforced for one member in each of these pairs. The reinforcement was given after the correct responses to the selected stimuli, irrespective of whether *S* had previously learned the correct response or had to read it upon exposure in the right window.

Initial criterion requirements consisted of two trial blocks of perfect anticipations, and then one more perfect block after a 5-min. rest period. The first test trial came next, followed by learning trials to one more perfect block, and then the second test trial. Figures A′ and C′ (see Fig. 7) were shown during the testing, one (selected randomly) on the first test trial and the other on the second test trial. *S*s were instructed to respond with the "most appropriate" name during the testing, but no reinforcement was given regardless of the actual choice.

Results and Discussion

The frequencies of the various responses to the test stimuli are shown in Table 18. The heading "Good" includes names given to the test stimuli which were associated with reinforcement during the preceding learning. The heading Not "Good" has similar meaning. There is obviously no significant difference between the proportions of reinforced and nonreinforced responses within the category of appropriate responses. There is a slight increase in the number of Not "Good" responses relative to the number of "Good" responses from the first to the second test trials. This may be related to the experimental condition whereby no reinforcement was given during the testing. It is interesting to note that in six of the seven cases where an inappropriate response was given it consisted of a reinforced response.

TABLE 18

FREQUENCY OF OBTAINED TEST RESPONSES FOR THE *S*s IN EXPERIMENT VI

	Appropriate Responses		Inappropriate Responses	
	"Good"	Not "Good"	"Good"	Not "Good"
First Test Trial	12	13	4	0
Second Test Trial	11	15	2	1

The results of this experiment are consistent with the statement (in a footnote) by Jackson (1954), "A third experiment substituting simple verbal 'rewards' (e.g., 'good,' 'that's right') for monetary influences was attempted, but results from six subjects suggested that these were not significantly effective in modifying figure–ground perception" (p. 340).

Although the experimenter's response "Good" apparently had no effect on the choice of names in test responses to the ambiguous cues, the question arises as to whether or not it had any effect during the learning process. That is, was there faster learning of the figure–name pairs associated with "Good" as compared with the figure–name pairs for which the experimenter gave no response? Fig. 8 gives the learning curves of total number of incorrect anticipations per learning block over all *S*s for both "Good" and Not "Good" figures. There is little basis to believe that there is any difference in the efficiency of learning names reinforced by "Good" as compared with names not so reinforced.

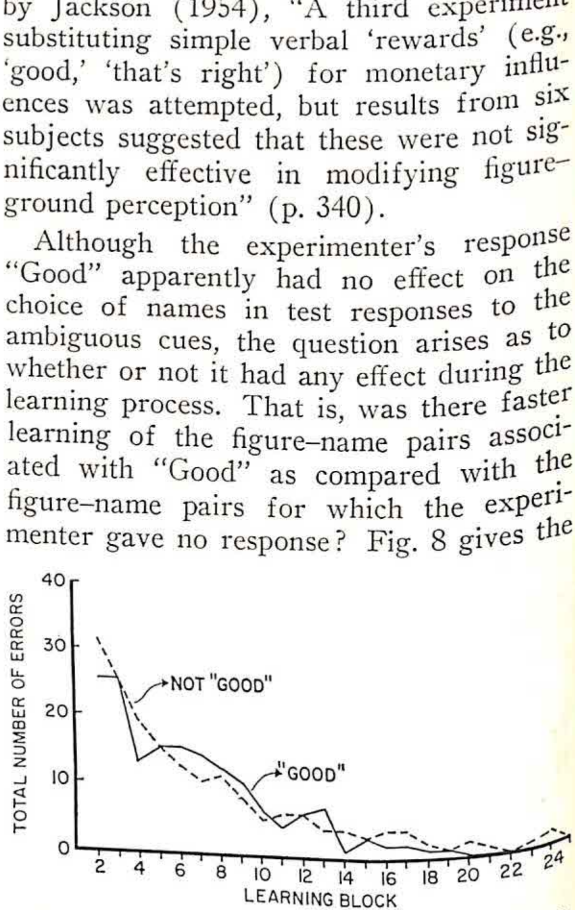


FIG. 8. Total errors over learning blocks for *S*s in Exp. VI.

Of course, the learning situation under consideration is quite different from the situation in the typical verbal conditioning experiment, where all response classes are immediately evident and learning is measured by an increase in the usage of one of these response classes at the expense of others.

The evidence, then, indicates little effect of the experimenter's response of "Good" upon the determination of responses to ambiguous cues or upon the facility with which responses in paired-associates learning are acquired. Accordingly, in the next experiment attention was focused upon a type of stimulation which seemed more likely to produce a learning effect.

EXPERIMENT VII

Electric shock was used as punishment in this experiment. The effects of punishment upon the determination of responses to ambiguous cues have been studied by Schafer and Murphy (1943), Proshansky and Murphy (1942), Rock and Fleck (1950), and Jackson (1954). But in these experiments any effects of punishment were completely confounded with those of reward, since one set of cues in every ambiguous configuration was previously always associated with the award of money while the other set of cues in the ambiguous configuration was associated with the withdrawal of money.

Smith and Hochberg (1954) used Schafer and Murphy type figures to investigate the effects of electric shock upon perceptual responses. Each face was exposed for $\frac{1}{3}$ sec. during learning trials, with one face in each reversible pair always associated with electric shock. As in previous studies, 63 presentations of the reversible figures were shown during test trials. There was a significantly greater number of nonshocked than shocked test responses over their entire group of 20 *S*s. Smith and Hochberg concluded that ". . . the 'punishment' (shock) procedure did affect the selection of the figure perceived" (p. 86). They considered their primary problem similar to that of Lazarus and McCleary (1951) who administered electric shock to *S*s in association with certain nonsense syllables. In a series of tachistoscopic test trials, Lazarus and McCleary then found that the average GSR for previously shocked syllables was higher than for nonshocked syllables despite incorrect recognition responses. Smith and Hochberg, contrary to Lazarus and McCleary, found explanations which did not involve "subception" or "perceptual defense" more consistent with their data.

In the Smith and Hochberg experiment, the electric shock accompanied presentation of the shocked face. This is not an entirely satisfactory procedure since the shock may serve as a cue, since it is a part of the total stimulus configuration, and thus function like any other cue during the test trials. Accordingly, in the present experiment the electric shock was presented subsequent to the naming response of the *S*. To be sure, this is contrary to the efforts of Lazarus and McCleary ". . . to avoid associating the shock with the subject's verbal report and thus influencing his response preference during the final test period" (p. 117). But in the present series of experiments interest was focused upon the responses made in recognition situations and the factors which influence these responses, and not in evaluation of such perceptual constructs as "autism," "subception," and "really see." Specifically, the question to which the experiment was directed was whether or not electric shock leads to a decrease in the tendency to use the response previously associated with a stimulus configuration when the configuration tends to evoke that response as well as others.

In order to assess the effects of differences in the intensity and method of administration of the electric shock, three groups of *S*s were run. For one group, a fixed shock was used for all *S*s; for the second group, the intensity of the shock administered to a given *S* depended on his verbal reports; and, for the third group, an attempt was made to persuade each *S* to take a higher intensity of shock than he would have received if he were in the second group.

Method

Subjects. All *S*s were students of introductory psychology at Indiana University. A total of 92 *S*s were run, 21 in Group VII-FS, 29 in Group VII-VS1, and 42 in Group VII-VS2.

Apparatus and materials. The apparatus for the controlled exposure of paired associates, as described in Exp. I, was used. The electric shock was administered by a constant current device, calibrated in milliamperes. The maximum possible shock was 10 milliamperes.

The learning figures and test stimuli used in this experiment may be seen in Fig. 7; the nonsense syllables RUX, VED, JOM, and NAF were paired with them. The stimuli lettered A and C in Fig. 7 were shown during the test trials.

Procedure. The *S*s in all groups were presented with the figures and nonsense syllables as paired associates during learning. A learning block consisted of the presentation of each figure–name pair once. Electric shock was administered to all *S*s during the learning trials in association with either A_1 or A_2 and its response, and also in association with either C_1 or C_2 and its response. The particular member within each of these pairs which was shocked was selected randomly over *S*s.

All *S*s were told that "a very mild electric shock" would be administered and were given a chance to leave if they objected to the shock for either medical or personal reasons. An effort was made to dissuade from leaving those *S*s whose objections seemed based on anxieties very little above the general public's fear of electricity. Nevertheless, about 5% of the total available set of *S*s were lost at this stage.

All *S*s in Group VII-FS were given a shock of six milliamperes. The intensity of the shock for the *S*s in Group VII-VS1 was set at that level which produced the verbal report indicating a pain threshold. Thus, after the electrodes were adjusted on the wrists of the *S*s and before the experiment was started, the following instructions were read:

> I would like to establish the intensity of the electric shock that will be administered during the course of the experiment. As I gradually increase the current intensity through a series of trials, I would like you to tell me the level at which you find the shock "just painful." Before each trial, I will say, "Ready" and then administer the electric shock. Remember, I want you to tell me when the shock is "just painful."

The first level of shock administered was zero milliamperes, the second 2 milliamperes, the third 4 milliamperes, and so on through 6, 7, 8, 9, and 10 milliamperes. The series of increases was stopped either at 10 milliamperes or whenever the *S* reported pain.

An attempt was made to get the *S*s in Group VII-VS2 to accept a current somewhat above the "just painful" level for the *S*s of Group VII-VS1. The preliminary instructions, after the electrodes were fastened, for these *S*s were like those reported above for the *S*s in Group VII-VS1. However, the first shock administered was two milliamperes, and this was increased by two miliampere steps to 10 milliamperes. Additional comments were made during the succession of increases in order to make each *S* a little less self-conscious during the process. Thus, prior to the two milliampere shock, *E* said, "Here is the lowest shock level. Usually people do not feel anything at the first shock intensity. Ready?" And after this shock was administered, *E* asked, "Did you feel anything?" Such comments as, "Now the second level," "Feel anything?" and "How was that?" were made during the subsequent course of intensity determination. When a level of current was reached after which the *S* reported that he could take no more, or that the shock was too painful, or that he wanted to stop, *E* said, "May I try it just a little higher? If it is too high I would be glad to turn it back down." After obtaining permission, the current was increased by one milliampere. Four *S*s refused to try the increase, and only three insisted on the reduction after the increase was made.

Learning proceeded to a criterion of two perfect blocks, a 5-min. rest period, and then another block of perfect anticipations. Two test trials were then given with the trials necessary for one more perfect block coming between them. On the test trials both Stimuli A and C were shown (see Fig. 7), in random order, to all *S*s under instructions to respond with the most appropriate name.

At the conclusion of the experiment all *S*s were asked to state why they gave their particular responses.

Results and Discussion

The mean shock intensities for the *S*s in Groups VII-VS1 and VII-VS2 were 8.2 and 9.1 milliamperes, respectively. The efforts to persuade the *S*s in Group VII-VS2 to accept higher levels of shock were apparently effective.

The response frequencies to the test stimuli in terms of previous association with shock or not-shock may be seen in Table 19. The responses are grouped into "appropriate" and "inappropriate" categories. It is perhaps most useful to consider initially the consistent pattern of alteration over all subject groups from first to second test response. In each case there is a clear shift toward more shocked responses and fewer not-shocked responses. Since no shock was administered during the test trials, it would seem reasonable to assume that this absence is the basis for the altera-

TABLE 19

Frequency of Obtained Test Responses for the *S*s in Exp. VII

	Appropriate Responses		Inappropriate Responses	
	Shocked	Not Shocked	Shocked	Not Shocked
Gp. VII–FS				
First Test	9	10	1	1
Second Test	12	8	0	1
Gp. VII–VS1				
First Test	5	21	2	1
Second Test	11	18	0	0
Gp. VII–VS2				
First Test	14	23	3	2
Second Test	19	18	4	1

tion. This is consistent with the findings of Exp. VI.

Accordingly, primary attention was paid to the responses to the first test stimulus for each *S* in determining the effects of the previous shock upon responses to the ambiguous cues. It is clear that the shock of fixed intensity had no significant effect upon the responses to the first test stimuli by the *S*s of Group VII-FS. The deviation from equal response freqencies for the *S*s in Group VII-VS1, on the other hand, is significant at the .01 level (using the binomial test). In contrast, the deviation from chance responding for the *S*s in Group VII-VS2 is not significant (using the χ^2 approximation since the cell frequencies are sufficiently large). To be sure the discrepancy goes in the right direction in that more *S*s gave a not-shocked response than a shocked response. But the ratio in that direction was only about 1.5 to 1 where it was more than 4 to 1 for Group VII-VS1. Assuming that the difference between groups is not merely chance, the evidence thus indicates that the extra efforts of *E* to persuade the *S*s to accept a higher level of shock decreased rather than increased the effect of the shock.

How can one account for this seemingly odd result? First, it must be remembered that seven *S*s did not accept the one milliampere increase in shock intensity during the shock adjusting trials. Presumably these seven *S*s were among those for whom the shock was most upsetting. To a certain extent at least, then, there may have been a tendency for the increase in average shock intensity for Group VII-VS2 as compared with Group VII-VS1 (9.1 vs. 8.2 milliamperes) to be associated with *S*s who were least disturbed by shock.

One other factor must also be considered. This is related to the finding that differences in the personalities of experimenters can produce differential learning effects in verbal conditioning (Binder et al., 1958). While the same *E* ran all three groups of *S*s in the present case, we note the slight increase in participation of *E* with Group VII-VS2 as compared with Group VII-VS1. While the goal was to make each *S* less self-conscious of the shock during the preliminary attempt to fix the maximum possible intensity, *E*'s increased responsiveness may have had a kind of psychotherapeutic result in ameliorating the painful effects of the shock.

The comments of the *S*s at the conclusion of the experiment as to their reasons for giving their responses are of some incidental interest. These comments indicate that the test responses were generally not "interpretive" in the sense that some writers on perception use the term: that is, there was no concomitant implicit relating of shock and response alternatives prior to naming the ambiguous stimulus. Thirteen of the total of 92 *S*s in Exp. VII related their choice of test responses to the previous shock during learning, and, of these, five actually gave shocked responses to the ambiguous cues. Among the stated reasons for test response choice are the following quotes, which are slightly paraphrased—the numbers and symbols following each quote indicate whether the *S* who made the particular statement gave two shocked (2S), two not-shocked (2NS), or one shocked and one not-shocked (1S, 1NS) responses:

Remembered them more strongly (2S)
First two learned (2NS)
Liked those stimulus words better, sounded prettier (2NS)
Paid no attention to the shock (1S, 1NS)
Looked more alike (1S, 1NS)
Easier to remember, especially JOM (1S, 1NS)

Didn't want to get shocked again (2NS)
RUX always shocked and used to saying it (2NS)
First thought of (1S, 1NS)
Both circles, means nothing, so like a blank space (2NS)
Knew it best, first one should have been JOM, I guess, because of jagged line (2NS)
Just stuck in my mind (2S)
Both were circles and therefore easier to remember (1S, 1NS)
Because these were ones shocked (2S)

Most *S*s stated that there was no reason for their particular choices, and next in order of frequency were statements indicating ease of remembering and priority of learning. Interestingly enough, and in accord with the findings in the area of verbal conditioning, most *S*s indicated during the conversation following the experiment that they were not aware of the particular pairs that were punished.

The results of this experiment, then, indicate that, when *S*s are confronted with an ambiguous situation in which a previously shocked and a previously unshocked name are appropriate, they will tend to respond with the unshocked name. This conclusion stems not only from the significant difference for Group VII-VS1, but also from the consistency of results (though insignificant) over all groups.

The effects of the shock upon acquisition during the paired associates learning trials may be seen in Figs. 9, 10, and 11. (The curves for Group VII-VS2 are somewhat curtailed since one individual made one error to a shocked stimulus in Block 19 and two errors to a not-shocked stimulus in Block 21.) Particularly noticeable is the consistent tendency for more errors to be associated with the shocked than the not-shocked stimuli from about the third to the eighth block. After approximately the eighth block (but varying between the seventh and tenth blocks for the different groups) there does not seem to be any consistent pattern. Despite the general consistency of these intervals in which more errors occur to the shocked than the not-shocked stimuli, there are obvious differences among the intervals. And these are, interestingly, related to the differences among the groups in the ratio between shocked and not-shocked responses to the test stimuli. Thus, the length of the interval for Group VII-VS1 is approximately seven blocks; that for VII-VS2, six blocks; and that for VII-FS, four blocks.

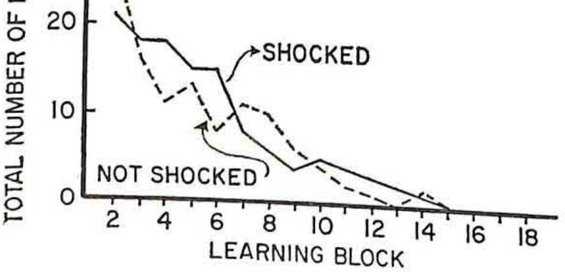


FIG. 9. Total errors over learning blocks for *S*s in Group VII-FS.

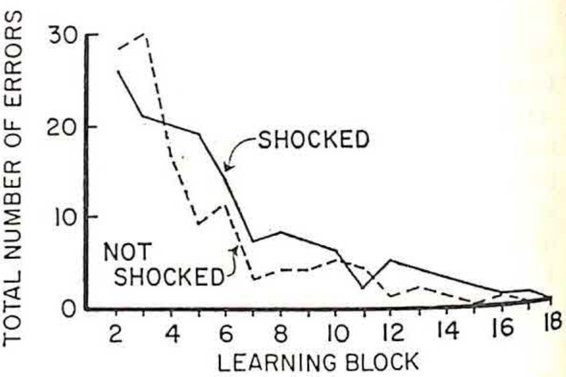


FIG. 10. Total errors over learning blocks for *S*s in Group VII-VS1.

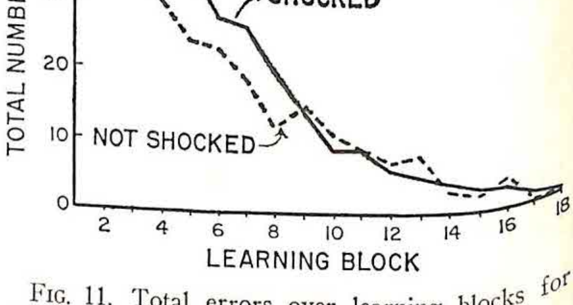


FIG. 11. Total errors over learning blocks for *S*s in Group VII-VS2.

A final comparison of some interest involves the proportion of errors, subsequent to the initial criterion of two perfect blocks, which were associated with shocked and not-shocked stimuli, respectively. These proportions and the frequencies which they represent may be seen in Table 20. The over-all difference of 38 shocked as against 19 not-shocked errors is significant at the .02 level (using χ^2).

Thus, the electrical shock seemed to have some effect in interfering with the learning of the stimulus–response connections with which it was associated, as well as leading to a difference in response tendencies to ambiguous cues. There is no compelling evidence indicating that the testing differences are accountable solely on the basis of the learning differences, nor the contrary, for that matter.

OVER-ALL DISCUSSION AND CONCLUSIONS

An immediate question that arises with respect to the frequency results is whether they are accountable on the basis of the development of response biases, rather than on the basis of recognition responses associated with particular stimuli in the previous learning trials. In other words, would comparable results have been obtained if blank cards had been shown during the test trials in place of the stimuli of reduced cues? Goldiamond and Hawkins (1958), for example, have shown that, under certain restrictive conditions, the relationship between word frequency and tachistoscopic threshold can be explained in terms of response biases. In their experiment, after a series of training trials involving the presentation of nonsense syllables at differential frequencies, *S*s were shown blank flashes at tachistoscopic speeds under instructions to respond with one of the previously learned syllables on each trial. A simulated method of limits was used. Goldiamond and Hawkins suggested that their results indicating the operation of response biases are generalizable to other types of studies establishing relationships between frequency and recognition thresholds.

TABLE 20

PROPORTIONS OF ERRORS AFTER INITIAL CRITERION IN EXP. VII

	Shocked	Not Shocked
Gp. VII–FS	70% (7)	30% (3)
Gp. VII–VS1	66.7% (6)	33.3% (3)
Gp. VII–VS2	65.8% (25)	34.2% (13)
Total	66.7% (38)	33.3% (19)

Note.—Values in parentheses give actual numbers.

There is evidence to indicate that in the present experiments the important frequency variable in determining test responses is the relative frequency of occurrence of a given stimulus under conditions in which the response in question is appropriate, and not merely the relative frequency of that response during the learning trials. First, there are the very few inappropriate responses during the test trials, which indicates that the *S*s were certainly attentive to the test stimuli. Second, the situations in which the relationship between relative frequencies during learning and relative frequencies among responses to ambiguous cues did not hold provide further evidence. Consider the *S*s in Group II-C and those in Exp. III: the relationship held with the latter but not with the former. Since the only difference in experimental conditions between the two groups was in the relative frequencies during learning (3:1:1:1 vs. 3:2:2:1), it is difficult indeed to see how the discrepancy in results between the two groups of *S*s can be explained by response biases, without stimulus considerations. (Stimulus considerations were of course essential to the explanation offered earlier for the discrepancy.) In the case of the Groups I-U and II-AB to Test Stimulus A, the discrepancy in results was associated with the presence and absence of a supplementary set of figures and names (the C_i's) during learning. The source of the difficulty seemed to be a stimulus confusion between A_1 and C_4. Moreover, during learning the frequencies of the appropriate responses to the test stimuli were identical for the two groups, so response frequency per se is certainly not adequate to explain the discrepancy. The C_i figures and their names are somehow involved, and it is not clear how their presence or absence can affect response biases.

The third line of evidence indicating that the results of the analysis of frequency in recognition are not explainable in terms of response biases stems from considerations of the actual frequencies of the various responses during the learning trials of the *S*s in Exp. III. These particular *S*s were chosen since all possible test responses were appropriate for them and they showed almost perfect correspondence between learning and test relative frequencies. Table 21 shows the number of erroneous responses, for each block of the paired-associates trials, consisting of names occurring three times, twice, and once per block, respectively. When an erroneous anticipation was of a form which was not clearly classifiable in terms of frequency, it was not included in the analysis. Examples of responses so excluded are GAF (derived from GEP and NAF), TUX (derived from TUZ and RUX), and VAL (derived from VED and KAL). In about 25% of the cases of errors in anticipation, the *S*s gave no responses prior to the time the name appeared in the right window; these could of course not be considered in this tabulation.

TABLE 21

Classifiable Erroneous Anticipations during Learning Blocks of *S*s in Exp. III

Block	Frequency of S–R Pair		
	3	2	1
2	14	65	31
3	13	45	32
4	10	29	15
5	3	19	12
6	3	11	6
7	3	9	3
8	0	3	1
9	1	3	3
10	0	2	0
11	0	1	0
12	0	1	0
Total, Blocks 2–4	37	139	78
	(0.48)	(1.78)	(1.00)
Total, Blocks 5–12	10	49	25
	(0.40)	(1.96)	(1.00)

If the differential frequencies of test results are explainable as response biases, these biases have surely developed during the learning trials and should be evident in the response tendencies during learning. Thus, whenever an erroneous anticipation occurs, one would expect that it is more likely to be a high frequency than a low frequency name and that this tendency will increase over the learning trials. The former is difficult to evaluate because of the complication that an *S* who responds to a stimulus figure in accord with a response bias and independently of the actual stimulus is more likely to be correct when he gives a high frequency rather than a low frequency name. But, despite this complication, if response biases are operative, the trend over learning blocks should be apparent in the proportions of erroneous anticipations for the S–R pairs of differing frequency. Table 21 shows no such trend. The numbers have been summed over blocks two to four as well as over all blocks from five on to facilitate this evaluation further. These totals are also included in Table 21; the parenthetical values are the totals in terms of the number in the right hand column as a unit. There is surely no evidence for the operation of response biases in these data.

As a final indication of the importance of stimulus factors in determining test responses we have the data on selective sampling. The *S*s in Group IV-C′-NE and those in IV-C-NE differ markedly in their proportions of various test responses, yet both groups gave the same average numbers of all of the various responses during learning. Similar comparisons may be made among other groups in Exp. IV that differed in either specific test stimulus or errors during criterion trials, but not in response frequencies during the learning trials.

Turning now to another aspect of the frequency analyses of responses to ambiguous cues: it is perhaps not unreasonable to assume on the basis of the data from groups that an individual's probability of a given recognition response to a set of ambiguous cues will be equal to the relative frequency of his previous S–R experiences with the named stimulus object.

The findings are especially interesting when consideration is given to the deductions and empirical evidence stemming from statistical learning theory. Estes and Straughan (1954) used an experimental situation involving a given set of stimulus elements and two response classes, with respective probabilities π and 1-π of being reinforced. They demonstrated mathematically and empirically that asymptotic probability of response is equal to probability of reinforcement under their specified conditions. Burke and Estes (1957) later showed that statistical learning theory predicts that response probability approaches reinforcement probability regardless of the distribution of sampling parameters (Estes and Straughan had assumed them equal for all elements). Finally, Estes (1957b) has presented the most general asymptotic matching formulation, showing that cumulative response proportion tends toward cumulative reinforcement proportion no matter how the latter varies over a series of trials. Moreover, in this presentation Estes allowed any number of alternative response classes.

Since the *S*s in the present experiments were required to correct all wrong anticipations upon exposure of the response syllables, the present experiments satisfied the necessary conditions of statistical learning theory that each possible response must have an associated reinforcement which, when present on a given trial, assures that the response will terminate the trial. Using the terminology of the statistical learning theorists (see Estes, 1957b, for example), the nonsense names appearing in the right windows of the present experiments correspond to the set of reinforcing events (E_1, E_2, . . . , E_r), the naming responses of the *S*s to the set of alternative responses (A_1, A_2, . . . , A_r), and the relative frequency of occurrence per trial block for the various stimulus–response pairs to the reinforcement probabilities (π_1, π_2, . . . , π_r).

Working within the framework of statistical learning theory, Peterson (1956) used an anticipatory technique to establish two verbal response hierarchies by associating responses consisting of English nouns with pairs of nonsense syllables used as stimuli. One of the syllables in each nonsense stimulus pair was common to another stimulus pair and thus conditioned to two noun responses, while the other syllable was unique and conditioned to a single response. He found, in word association test trials involving the presentation of the ambiguous nonsense syllables (those common to two noun responses), that the observed relative frequency of response was equal to reinforcement probability (as corrected for the slight dissimilarity between learning and testing situations).

Schoeffler (1954) used test patterns composed partly of stimuli to which a given response had been conditioned, partly of stimuli to which a competing response had been conditioned, and partly of stimuli to which neither the given nor the competing response had been conditioned. His stimuli consisted of light patterns selected from a group of 24 lights, and his responses of movements of a lever to the right or left. He found that the probability of a response to a test pattern is equal to the proportion of stimulus elements in the test set that are conditioned to the response. It should be noted that Schoeffler produced ambiguity in his test stimuli by presentations containing too many cues.

The finding in the present experiments of a close correspondence between relative frequencies of presentation during learning (reinforcement probabilities) and relative frequencies of test responses provides general support for the asymptotic expectations based on statistical learning theory. However, it must be noted for the experiment involving the faces (Exp. V) that, since all cues were given 100% reinforcement, the analyses of Estes and Straughan (1954), etc. do not lead necessarily to the prediction of matching of test response and learning relative frequencies—that is, if one assumes a high degree of selective sampling such that the common elements have a comparatively low probability of being sampled. Deviations from matching would be expected to be in the direction of equal frequencies between the two alternatives (which is in accord with the results shown in Table 15). But the status of the common elements in this particular experiment is not at all clear because of their central, attention-getting characteristics.

Let us consider the other general result, namely, that *S*s who learn to make discriminating responses indicating recognition on the basis of configurations can make accurate discriminations using parts of these configurations. In only one of the six comparisons in Exp. I did the frequencies of appropriate and inappropriate responses to the test stimuli differ significantly from the frequencies expected purely as a result of the *S*s not having achieved perfect learning levels. Moreover, the results of the subsequent experiments seemed to confirm this general result.

These results of appropriate and inappropriate response frequency have important implications for conceptualization in statistical learning theory. They indicate the limitations of the approach involving component elements where stimulus events or elements are equated with cues and no allowance is made for the interaction of cues or the discriminations based upon configurations. Thus, a component approach would, in the present experiments, involve identifying various cues (circle or triangle, wavy or straight line at top, slanted line in right or left corner as in Fig. 1) or their parts as stimulus elements without assigning stimulus elements to patterning effects (such as a triangle with a wavy line above it). The stimulus elements associated with the pattern would be the stimulus elements identified separately with the components of the pattern and nothing more.

Estes (1957a) has questioned the adequacy of the component approach in statistical learning theory and pointed out the improvement in predictions resulting from considerations of pattern or configuration. In terms of the present experiment, statistical analysis of the stimulating situation based upon components leads to the prediction of many more inappropriate responses than were actually obtained. Let us demonstrate this under the assumptions that all stimulus elements had been perfectly conditioned prior to test trials and that the *S*s had learned to sample selectively early in learning. The present argument would be even stronger (though more difficult to present) without these assumptions.

Consider the stimulus elements associated with exposure of Test Figure A (see Fig. 1) in accord with a component-element approach, using Group I-E as an example. One-fourth of the stimulus events sampled in observing the triangle have been connected to the name of Figure A_1. One-fourth of the events have been connected to the name of A_2, another fourth to the name of C_3, and the final fourth to C_4. Similarly, the stimulus events associated with the line above the triangle in A have been evenly conditioned among the responses to A_1, A_2, B_1, and B_2. There would seem to be no basis for assuming that events connected with the names of A_1 and A_2 are any more likely to be sampled on a given testing trial than other events. Thus, since the probability of a given response is considered equal to the proportion of stimulus events in the population which are conditioned to the response, the probability of the name of A_1 to A is $\frac{1}{4}$; the name of A_2, $\frac{1}{4}$; either the name of B_1 or the name of C_3, $\frac{1}{4}$; and either the name of B_2 or the name of C_4, $\frac{1}{4}$. This leads to the expectation of 50% appropriate and 50% inappropriate responses to Test Stimulus A. A similar analysis may be made for Test Stimulus B for the *S*s in Group I-E. A comparable analysis for the *S*s in Group I-U involves some complications due to the unequal trial-block frequencies, but can be expected to produce similar results.

Thus, some of the results which would be predicted on the basis of a stimulus-component approach to statistical learning theory are at considerable variance with the obtained results. This clearly indicates the desirability of component interaction or, in other words, patterning considerations.

The same data also indicate the dangers in going too far in configurational emphasis. The findings on appropriate and inappropriate responses raise questions for the gestalt interpretation of the phenomena of recognition. Thus, Koffka (1935) states:

> In applying the law of similarity to the selection of a trace by a process one has to exercise great care. One must bear in mind that similarity must exist between, and therefore be defined in terms of, the process and the trace; hence similarity or partial identity between the two sets of *stimuli* which give rise to the formation of the trace and the process now occurring is *not* an adequate criterion for the application of the law. And therefore any explanation like the following one is essentially wrong: If an organism has reacted in a certain way to the stimulus complex A B C D E F, it will later respond in the same way to the recurrence of a part of this complex, say B C D. . . . We know that the response to a stimulus complex is not the sum of all responses to its individual components, but an organized pattern in which each part depends upon the organization of the whole (p. 601).

Koffka does, however, allow set or attitude much influence in producing communication between "process and trace." As an example of this he shows how, under the influence of proper set, a previously experienced figure is seen as part of the stimulus complex of a more complicated figure. Perhaps a gestalt psychologist might argue that, since the *S*s had to make discriminations on the basis of all cues in the complex, this produced an attitude which enabled them to discern and utilize the relationship between stimulus complex and stimulus part.

In summary, the results of the experiments dealing with learning frequency provide very convincing evidence of a direct relationship between learning frequency and perceptual response frequency. Moreover, these results provide a set of conditions under which the relationship breaks down. In addition, the matching of learning and recognition frequencies was useful in the investigation of selective attention in perceptual learning. It was found in this experiment that *S*s who maintain errorless responding after attaining initial criterion levels tend to respond on test trials in a manner which indicates selective attention to cues during the process of learning perceptual responses. For these individuals, common and discriminative cues among the stimuli during learning differ in their availability for association with the various responses.

Surely it is clear that a good deal of additional evidence must be accumulated before the phenomena of selective attention in recognition become clarified. Further work might also very profitably be aimed at the precise specification of the conditions under which the relationship between relative frequency of occurrence during learning and relative frequency of perceptual response does and does not hold. This would seem to have important implications not only for an analysis of such ambiguous stimulus phenomena as tachistoscopic and projective presentations, but also for statistical learning theory. Indeed, the results obtained here provide at least a partial bridge between the two areas.

The results of the experiments involving reward and punishment during learning are not as clear or compelling. Surely the indications are that shock in association with a recognition response during learning makes that response less likely to occur than another appropriate response for an ambiguous cue configuration. The experimenter's response of "Good" apparently has no comparable influence in determining responses to ambiguous cues.

An important consideration in studies on the effects of reward and punishment upon perception is the point at which the reinforcement occurs during the learning trials. Proshansky and Murphy (1942) apparently presented their rewards and punishments approximately simultaneously with their various stimuli, although the exact timing is not made clear in their article. Moreover, the *S*s made no overt responses during the training period. Schafer and Murphy (1943) report that reward and punishment came after *S* had repeated aloud the name which the experimenter had presented orally just prior to exposure of the face. The sequence in the case of the experiments of Rock and Fleck (1950) and Jackson (1954) was as follows: name told by experimenter, stimulus presented, *S* repeated name, and, finally, reward or punishment given. Apparently, Solley and Sommer (1957), however, presented the stimulus first, then pronounced the name for the *S*, and administered reward at some time after the face appeared. No specific mention is made of whether the *S*s did or did not pronounce the names during the training series, but miscellaneous comments in their article give the impression that the *S*s did not do so. Smith and Hochberg (1954) administered shock simultaneously with the presentation of the stimuli. For the first 12 trials, the name was given to each *S* prior to each presentation, and thereafter *S* called out the name of every figure after it was shown.

The administration of reward and/or punishment after exposure of the stimulus, but prior to the response of the *S* (whether overt or not), does not seem reasonable if one is interested in the reinforcement aspects, rather than the cue aspects, of the reward and/or punishment. This is most obvious in the case where either reward or punishment is associated with one stimulus and nothing with the other stimulus of an ambiguous team. The finding of Smith and Hochberg, for example, of significantly more unshocked than shocked responses to the ambiguous test stimuli is explainable in terms of the absence of the electric shock as a cue for one of the two possible responses. One further difficulty in many of these studies is that the relationship between reward and punishment, on the one hand, and the conditions of learning, on the other, is obfuscated because the correct responses were given either prior to or immediately after exposure of the stimuli, giving the *S*s no chance for anticipation and subsequent reinforcement or correction. (There is evidence, as one might expect, that the *S*s in these experiments showed very inefficient learning, as demonstrated by enormous numbers of erroneous test responses. This is discussed by Rock and Fleck, 1950).

But associating reinforcement with response in paired-associates learning, as in the present experiments, is objectionable to those who argue that such a procedure may affect the response but not the perception. To illustrate this position, in the case of their experimental situation involving shock, Smith and Hochberg (1954) distinguish between "autism" and "inhibition of the verbal 'naming' response," and indicate that their procedure produced the autistic effect without verbal inhibition.

The problem of differentiating what is perceptual and what is purely verbal in a perceptual-type experiment has been particularly vexing to research workers in this area. One attempt at a solution to this problem has involved complete operationalism, where perception is defined in terms of the experimental conditions and recognition as the association of a correct name with a given pattern. As Estes (1959) has pointed out, the difficulty with this operational viewpoint ". . . is that, as always when one puts on blinkers, we would be in danger of failing entirely to see important aspects of the problem" (in press). However, investigators like Murphy, Schafer, Proshansky, Pastore, Smith, Jackson, Hochberg, etc. have surely not been guilty of moving too far in the direction of operationalism. In fact, some of their statements seem to indicate that a few have given perception or the "really see" process a scientific reality independent of the observations which provide its anchoring. For example, in his argument that the experimental tasks of Proshansky and Murphy (1942) were judgmental and not perceptual in nature, Pastore (1949) states:

> In order to subsume the experiment under the category of perception, the authors must demonstrate that the visual experience of the subjects is actually changed; the rewarded lines must be actually magnified in the visual experience of the subjects (p. 470).

But all investigators, regardless of theoretical persuasion, do evaluate perception experimentally by means of some recognition response under specified conditions. To be sure, other observations are sometimes made and reported to provide anchoring for a position inferring perceptual and not purely verbal processes. Thus, in support of their viewpoint that the *S*s "actually perceived" in accord with the training, Proshansky and Murphy (1942) point out that their *S*s frequently complained of the difficulty of perceiving the stimuli, that the *S*s displayed a good deal of interest in the stimuli, and that the *S*s responded carefully and not with undue haste. Similarly, in arguing for a genuine reorganization of the perceptual field on the part of their children, Solley and Sommer (1957) state that the *S*s reported the reward-associated figure as phenomenally nearer and as having a dark line around it, producing contour emphasis.

The fact of the matter is that so little is actually known about the perceptual process and its indicators that many of the discussions on the perception vs. naming problem are purely in the realm of polemics. Current theories on the factors involved in making differential responses to stimulus configurations are quite inadequate and desperately in need of clarifying data. All investigators involved in perceptual-type research are interested in the responses of *S*s under certain rather loosely defined conditions. And, from the viewpoint of emphasis received in the literature, surely the naming or identifying response is most important. With the development and maturation of a theory, perhaps other types of responses will be equally well investigated for differentiation among higher-order constructs, but the groundwork must be firmly laid in terms of the more basic data. The experiment of Neisser (1954) is illustrative of the process of gradual elaboration of a relatively narrow perceptual concept upon the base of a well-established core. After making a rough distinction between what an *S* sees and what he says, he attempts to show that familiarity with words facilitates "seeing rather than saying the words" in tachistoscopic presentations. Defining his concepts in a very restricted fashion, he demonstrated that prior experience with words lowered the tachistoscopic thresholds for these words but not their homonyms.

The present research project has been concerned broadly with those factors in the process of learning differential responses to stimulus configurations which affect naming responses to ambiguous cue sets. In accord with most of the research in the area of perceptual processes, the test response variable studied was naming or identifying. Thus, returning to the point of departure for this particular discussion, the question of whether the response or the perception is affected by reinforcement applied during acquisition of the naming responses is not particularly meaningful at this rather crude stage of perceptual theory.

One surely gets the over-all impression from these results *in toto* of the greater significance of frequency of occurrence than of differential reward and punishment in influencing responses to ambiguous cues. The most extreme ratio of one test response to the other possible response was 21 to 5 (see Table 19) in the case of the experiments on reward and punishment, but that order of ratio was obtained repeatedly in the frequency experiments when one learning stimulus appeared four times as often as its testmate (see Tables 3, 7, and 17). Beyond that there are the precise predictions possible with learning frequency as the independent variable and the correspondence of results with the theorems of a well-formulated theory, advantages which do indeed seem to be a long way off in the case of reward and punishment.

SUMMARY

The over-all aim of this research was to investigate the effects of variations in the conditions of learning upon responses to ambiguous stimuli. The learning consisted of trial blocks of figure–name pairs learned to a criterion by the method of paired associates. Test trials, in which the *S*s were presented with cues common to previously learned figure groups, followed the learning trials. The *S*s were instructed to respond to the ambiguous cues with the one of the previously learned names which seemed most appropriate.

The learning variable of interest in the first set of experiments was relative frequency of occurrence. Eight groups of *S*s were run in five experiments. For seven of these groups the learning figures were composed of arbitrarily chosen cues and consequently meaningless. The learning figures for these groups contained common as well as differentiating stimuli, and the test stimuli consisted of only the common cues. For the remaining group the learning figures consisted of line profiles of human faces with the lines continued to form semicircles. The profiles were constructed in pairs having common contours, and the test stimuli consisted of these common contours in a full circle.

Among the groups whose learning figures were completely meaningless, two groups were each shown a total of eight figures per block. For one group the figures occurred at equal frequencies per block, while for the other group the frequencies per block for the various figures were (2:1) : (4:1) : (3:1:1:1) (where the parentheses show the frequencies of figures having common cues on a test trial). Three other groups in the category of those shown meaningless learning figures were shown four figures, at frequencies, respectively, of (2:1) : (4:1), (3:1:1:1), and (3:2:2:1).

The responses to the test stimuli were either consistent with the test cues, in the sense that the response names were of learning stimuli which contained these cues, or the responses were names of learning stimuli having cues which conflicted with the test cues. A hypothesis stemming from a statistical model of visual recognition was supported by the finding that the relative frequencies among the responses consistent with the ambiguous testing cues tended to equal their relative frequencies of occurrence during learning. However, there were two test stimuli for which this matching did not hold: the (2:1) test for the (2:1) : (4:1) : (3:1:1:1) group and the test for the (3:1:1:1) group. Hypotheses for these discrepancies were offered and deductions from them confirmed. Further, a second major hypothesis that the learning of discriminatory responses results in the ability to identify subsets of the learned cues was generally confirmed. Recency as an explanation for the frequency findings was ruled out.

The remaining two of the seven groups shown nonsense learning figures were used to test the hypothesis that *S*s learn to ignore selectively the nondiscriminative cues among the stimuli in a multiple discrimination situation involving recognition responses. Use was made of the finding that *S*s tend to give recognition responses to test cues at the same relative frequencies as the occurrence of these cues as components of various previously seen learning figures. The test analyses included the naming responses of one group of *S*s to a stimulus containing the cues common to all learning figures, as well as the naming responses of another group to stimuli containing both discriminative and common cues. Both groups were shown the same four learning figures equally frequently per block for the first three blocks, and then at respective frequencies of (2:1:4:1) per block throughout the remaining learning trials.

Only those *S*s who made no errors in the learning trials subsequent to the initial criterion level were used in the test of the hypothesis of the selective use of cues. Among these *S*s, those who were shown the test stimuli containing both common and discriminative cues responded with the names of figures at relative frequencies not significantly different from the relative frequencies of occurrence of these figure–name pairs during the learning trials of differential frequency of presentation. However, those *S*s who were shown the test stimuli containing only common cues responded with names at relative frequencies which differed significantly from those expected if differential learning frequencies were matched, but did not differ significantly from the frequencies expected if the common cues were selectively ignored by the fourth learning block.

The group which had the profiles as learning figures were shown four of them at frequencies of (2:1) : (4:1). The matching between relative frequencies of occurrence during learning and relative frequencies among test responses held for their first response but not for their second; lack of independence resulting from the particular nature of the stimuli was hypothesized.

The learning variables of interest in the second set of experiments were reward and punishment. Four groups of *S*s were run and four learning figures were shown to each group. The learning figures were in pairs in the sense that reward or punishment was associated with one member of a given pair and no reward or punishment with the other. Reward or punishment followed response. Two test stimuli, one for the cues common to the members of each pair, were shown to the *S*s.

One group received "Good" as reward, while the other three groups received electric shock as punishment. These latter three groups differed principally in average intensity of shock. No significant difference was found between the number of responses which were previously rewarded by "Good" and the number not so rewarded. However, a tendency was found for *S*s to give test responses which were not previously shocked, at least in their first responses. The difference was significant for one group (medium shock intensity) and in the given direction for the other two.

REFERENCES

Arnoult, M. D. Familarity and recognition of nonsense shapes. *J. exp. Psychol.,* 1956, **51**, 269–276.

Atkinson, R. C. The observing response in discrimination learning. *USN ONR tech. Rep.,* 1959, No. 4.

Attneave, F. Psychological probability as a function of experienced frequency. *J. exp. Psychol.,* 1953, **46**, 81–86.

Binder, A. A statistical model for the process of visual recognition. *Psychol. Rev.,* 1955, **62**, 119–129.

Binder, A., McConnell, D., & Sjoholm, N. A. Verbal conditioning as a function of experimenter characteristics. *J. abnorm. soc. Psychol.,* 1957, **55**, 309–314.

Bruner, J. S. On perceptual readiness. *Psychol. Rev.,* 1957, **64**, 123–152.

Bruner, J. S., Miller, G. A., & Zimmerman, C. Discriminative skill and discriminative matching in perceptual recognition. *J. exp. Psychol.,* 1955, **49**, 187–192.

Burke, C. J., & Estes, W. K. A component model for stimulus variables in discrimination learning. *Psychometrika,* 1957, **22**, 133–145.

Chein, I., Lane, R., Murphy, G., Proshansky, H., & Schafer, R. Need as a determinant of perception: A reply to Pastore. *J. Psychol.,* 1951, **31**, 129–136.

Estes, W. K. Of models and men. *Amer. Psychologist,* 1957, **12**, 609–617. (a)

Estes, W. K. Theory of learning with constant, variable, or contingent probabilities of reinforcement. *Psychometrika,* 1957, **22**, 113–132. (b)

Estes, W. K. Statistical models for recall and recognition of stimulus patterns by human observers. *Proc. Interdisciplinary Conf. Self-Organizing Sys.,* 1959, in press.

Estes, W. K., & Burke, C. J. A theory of stimulus variability in learning. *Psychol. Rev.,* 1953, **60**, 276–286.

Estes, W. K., & Burke, C. J. Application of a statistical model to simple discrimination learning in human subjects. *J. exp. Psychol.,* 1955, **50**, 81–88.

Estes, W. K., & Straughan, J. H. Analysis of a verbal conditioning situation in terms of statistical learning theory. *J. exp. Psychol.,* 1954, **47**, 225–234.

Glaze, J. A. The association value of non-sense syllables. *J. genet. Psychol.,* 1928, **35**, 255–267.

Goldiamond, I., & Hawkins, W. F. Vexierversuch: The log relationship between word-frequency and recognition obtained in the absence of stimulus words. *J. exp. Psychol.,* 1958, **56**, 457–463.

Green, E. J. A simplified model for stimulus discrimination. *Psychol. Rev.,* 1958, **65**, 56–63.

Hake, H. W., & Eriksen, C. W. Role of response variables in recognition and identification of complex visual forms. *J. exp. Psychol.,* 1956, **52**, 235–243.

Howes, D. H., & Solomon, R. L. Visual duration threshold as a function of word-probability. *J. exp. Psychol.,* 1951, **41**, 401–410.

Jackson, D. N. A further examination of the role of autism in a visual figure-ground relationship. *J. Psychol.,* 1954, **38**, 339–357.

Koffka, K. *Principles of gestalt psychology.* New York: Harcourt, Brace, 1935.

Krasner, L. Studies of the conditioning of verbal behavior. *Psychol. Bull.,* 1958, **55**, 148–170.

La Berge, D. L., & Smith, Adrienne. Selective sampling in discrimination learning. *J. exp. Psychol.,* 1957, **54**, 423–430.

Lawrence, D. H. Acquired distinctiveness of cues: II. Selective association in a constant stimulus situation. *J. exp. Psychol.,* 1950, **40**, 175–188.

Lazarus, R. S., & McCleary, R. A. Autonomic discrimination without awareness: A study of subception. *Psychol. Rev.,* 1951, **58**, 113–122.

Levine, R., Chein, I., & Murphy, G. The relation of the intensity of a need to the amount of perceptual distortion: A preliminary report. *J. Psychol.,* 1942, **13**, 283–293.

Neisser, U. An experimental distinction between perceptual process and verbal response. *J. exp. Psychol.,* 1954, **47**, 399–402.

Pastore, N. Need as a determinant of perception. *J. Psychol.,* 1949, **28**, 457–475.

Peterson, L. R. Prediction of response in verbal habit hierarchies. *J. exp. Psychol.,* 1956, **51**, 249–252.

Proshansky, H., & Murphy, G. The effects of reward and punishment on perception. *J. Psychol.,* 1942, **13**, 295–305.

Restle, F. A theory of discrimination learning. *Psychol. Rev.,* 1955, **62**, 11–19.

Restle, F. Theory of selective learning with probable reinforcements. *Psychol. Rev.,* 1957, **64**, 182–191.

Restle, F. Toward a quantitative description of learning set data. *Psychol. Rev.,* 1958, **65**, 77–91.

Restle, F. Additivity of cues and transfer in discrimination of consonant clusters. *J. exp. Psychol.,* 1959, **57**, 9–14.

Rock, I., & Fleck, F. S. A re-examination of the effect of monetary reward and punishment on figure-ground perception. *J. exp. Psychol.,* 1950, **40**, 766–776.

Sanford, R. N. The effects of abstinence from food upon imaginal processes: A preliminary experiment. *J. Psychol.,* 1936, **2,** 129–136.

Sanford, R. N. The effects of abstinence from food upon imaginal processes: A further experiment. *J. Psychol.,* 1937, **3,** 145-159.

Schafer, R., & Murphy, G. The role of autism in a visual figure-ground relationship. *J. exp. Psychol.,* 1943, **32,** 335–343.

Schoeffler, M. S. Probability of responses to compounds of discriminated stimuli. *J. exp. Psychol.,* 1954, **48,** 323–329.

Smith, D. E. P., & Hochberg, J. E. The effect of "punishment" (electric shock) on figure-ground perception. *J. Psychol.,* 1954, **38,** 83–87.

Solley, C. M., & Sommer, R. Perceptual autism in children. *J. gen. Psychol.,* 1957, **56,** 3–11.

Solomon, R. L., & Postman, L. Frequency of usage as a determinant of recognition thresholds for words. *J. exp. Psychol.,* 1952, **43,** 195–201.

Taffel, C. Anxiety and the conditioning of verbal behavior. *J. abnorm. soc. Psychol.,* 1955, **51,** 496–501.

Taylor, Janet A. Meaning, frequency, and visual duration threshold. *J. exp. Psychol.,* 1958, **55,** 329–334.

Thorndike, E. L., & Lorge, I. *The teacher's word book of 30,000 words.* New York: Bureau of Publications, Teachers College, 1944.

Wyckoff, L. B., Jr. The role of observing responses in discrimination learning. Part I. *Psychol. Rev.,* 1952, **59,** 431–442.

(Received December 19, 1959)

Vol. 74, No. 10 Whole No. 497, 1960

Psychological Monographs: General and Applied

A RE-EVALUATION OF THE MEANINGFULNESS OF ALL POSSIBLE CVC TRIGRAMS

E. JAMES ARCHER

University of Wisconsin[1]

For too many years we psychologists who are interested in verbal learning have been using scales of meaningfulness of questionable validity. The most frequently referred to study of the association value of nonsense syllables has been that by Glaze (1928). Especially since only its results have been reported in a more readily available secondary source (Hilgard's chapter in Stevens, 1951), the shortcomings of the original study are usually unknown. The Glaze study had but 15 subjects (*S*s); many nonword, three-letter combinations were omitted from the listing; 29 items had two association values (which also raises a question of the precision of estimation of the single-valued items); and the manner of *S*'s responding may have placed some differential constraints on reporting associations since *S* had to verbalize them to the experimenter (*E*). Another and slightly more recent study has been used somewhat less (Krueger, 1934). This study had some advantages over Glaze's, but it also had some inadequacies. Although a total of 586 *S*s served in it, not all *S*s judged all syllables; instead, each association value was based on 200 *S*s. And although the *S*s wrote the syllables and their responses to each, they also had in view up to 49 of their previous responses. The considerably higher values reported by Krueger might have been due to the recognition of associations among the items in view and to the fact that *S*s were allowed considerably more time to respond than in either the Glaze or the present study. Furthermore, the first 100 responses were considered practice, and apparently the responses to these were discarded. As will be evident from the present study, the most serious fault was the failure to take into account the wide differences among individuals. If the final values do not represent the averages of all *S*s, the association values for individual items will not be comparable.

An additional reason for the present re-evaluation is that, since the above studies were made, an indeterminate number of so-called nonsense syllables have become acceptable words, slang expressions, and abbreviations.

This study is a re-evaluation of the meaningfulness of all possible three-letter combinations of the Roman alphabet of the form consonant-vowel-consonant, with the restrictions that the two consonants are different and neither is a y when y is the vowel. Because it seems incongruous to perpetuate the self-created paradox of talking about the meaningfulness of nonsense syllables, it is herewith proposed to call these combinations "trigrams," and in view of their form to further specify them as "CVC trigrams."

Since the determined meaningfulness might be a function of the verbal fluency of the *S*s and furthermore the two sexes might respond differently to a particular trigram, the fluency of the *S*s was measured and cognizance taken of the sex of the respondents.

PROCEDURE

Materials. There are 2,480 possible combinations of three letters of the Roman alphabet which meet the restrictions mentioned above. These trigrams were arranged in eight groups of 300 and one group of 80.

---

[1] The writer is indebted to Nancy L. Westby, who gave considerable assistance in the preparation of the materials and in the running of the subjects, and to the Numerical Analysis Laboratory of the University of Wisconsin, without whose facilities this study would have been impossible.

The orders were partially determined by Glaze's values. In order to prevent the establishment of artificial anchoring points, two restrictions were imposed on the sequence of presentation of the items: (*a*) where possible, successive items were of different association values and (*b*) successive items had no letters in common. Each of the nine sets of trigrams was made into a 16-mm. film strip with a blank space between successive items.

Method of presentation. The film strips were presented to groups of *S*s which varied in size from 12 to 24. The order of presentation of the film strips was determined by the rows of two different 9 × 9 latin squares, i.e., there were 18 sequences of film strips. Therefore, although the order of presentation of items was fixed within a film strip, the order of film strips was varied to prevent the establishment of artificial anchoring points. To obtain a check on reliability of judgment the first film strip was repeated after all others had been presented. Therefore *S* saw 2,780 or 2,560 items, depending upon the length of his first film strip.

The film strips were presented at a rate of 2 sec. per frame, except for the first 15 items in the first session. Since there was a blank space between successive trigrams (to encourage independent judgments), the usual time allowed for responding was 4 sec. The first 15 items in the first session were presented at a slower rate of 3 sec. per frame in order to assure that all *S*s could learn the required responding routine.

To aid *S* in keeping his place on the recording sheet, every fifteenth item was signaled. The scoring sheet had 30 15-item columns. In the blank space immediately preceding the first trigram a large "1" was printed in ink directly on the film. Between the fifteenth and sixteenth items a number "16" was likewise printed. Similarly, 31, 46, 61, etc. were printed. This method of printing was used to make the numbers obviously different from the trigrams. The numbers appeared as black on a grey background, while the trigrams were white on a grey background. If, in spite of the special numbering, *S* lost his place, his instructions were to "get in step" at the start of the next block of 15. After the complete showing of the film strip a specific section was repeated if needed. This was seldom necessary. The signaling system seemed adequate, and the *S*s proved quite adept at keeping their places.

An IBM form with 300 Y-N answer spaces was issued to each *S* before the start of each film strip. The sheet bore the number of the film strip he was about to see. Each *S* recorded his name, date, age, and sex on the sheet before the film strip was shown.

Subjects. A total of 335 University of Wisconsin students from several introductory psychology courses served in the study. For various reasons the data for 16 *S*s were discarded. The final, otherwise unselected, sample had 185 women and 134 men. Because the sequence of presentation of film strips proved significant, the final computations are based on 18 matched groups of 12 *S*s each ($N = 216$). In this final sample the mean ages of the men and women were 20.8 and 19.2 years, respectively.

Task. In order to maximize the *S*'s interest and motivation, the author (who also was a course instructor for many of the *S*s) gave a 15-min. explanation of the importance of the study, emphasizing why great care should be taken in making each response. Before the first film strip was presented, the *S*s were told that they were to pronounce each trigram to themselves and ask themselves the following questions: "Is it a word? Does it sound like a word? Does it remind me of a word? Can I use it in a sentence?" They were instructed to mark the scoring sheet in the "Yes" column if they answered "Yes" to any of the questions. If they could not answer "Yes" to any of the questions, they were to make a mark in the "No" column on the scoring sheet.

After each film strip the *S*s were given a 5-min. rest. Before each subsequent film strip the author again reminded the *S*s of the criteria for scoring an item as meaningful. In addition, they were urged to maintain their standards throughout the study. The *S*s saw 900 or 980 trigrams on each of three consecutive days. Sixteen of the groups were run in the evening; two were run in the afternoon.

Fluency test. After the showing of a film strip on the second day, the *S*s took a short verbal fluency test. The *S*s were instructed to write all of the words in the English language they could think of which began with the letter "s" and which were not proper nouns. They were allowed 5 min. at this task. Any questions about the task were deferred until the end of the experiment.

General conditions. The study was conducted in a classroom; and, to minimize interaction among *S*s, they were encouraged to spread out. The trigrams were projected onto a screen at the front of the room. For the evening sessions, only the overhead lights in the rear half of the room were used. This provided enough light to mark the scoring sheets, but the screen was dark enough to make the trigrams legible. The projected trigrams were about 5 in. high.

Either the author or Westby was in the room at all times during the showing of a film strip. The *E*s deliberately showed interest in the *S*s' work so as to maintain motivation and interest. After the first 1,000 or so trigrams, judging of meaningfulness got to be a less than exciting task. If, however, any *S* showed evidence of carelessness, persistently communicated with a neighbor, or was unable to keep his place on the recording sheets, his data were discarded. There were 16 such cases in all. In a few cases *S* would accidentally skip a column in his recording. The standing instructions were to continue the task and record the last column on the reverse side of the scoring sheet. The record was then transcribed correctly by the *E*s.

RESULTS AND DISCUSSION

Fluency. The correlation between number of words written on the fluency test and number of trigrams recorded as being meaningful (as defined previously) was computed for men and women separately. The obtained values were .128 (for 134 men) and .138 (for 185 women). Neither correlation is significantly greater than .000. If, however, the correlation is computed without regard to sex of the respondent, the value is 0.142 ($N = 319$), $.05 > p > .01$. The higher correlation is due to the increased range of fluency scores. The overall range of fluency scores was from 19 to 94 ($\sigma = 13.63$, $M = 53.01$).

Incidentally, there was a significant difference between the two sexes on the fluency test. Women wrote 55.4 words in 5 min., while men wrote only 49.7 ($t = 3.672$, 317 *df*, $p < .001$).

Meaningfulness. Before determining the percentage of *S*s reporting a trigram as meaningful, it was necessary to check on the effect of two potential variables: sequence of viewing and practice. Even though the film strips were made so as to minimize any anchoring effects, there was the possibility that the sequence of viewing the film strips could lead to differing levels of judgment of meaningfulness. In addition there was the possibility that the criterion of meaningfulness used by the *S* might change as a function of practice in the study.

In order to test the effects of sequence and practice, the data from 18 groups (corresponding to sequences of film strips) of 12 *S*s each were assembled. Although fluency was not highly correlated with the reporting of a trigram as meaningful, the groups were assembled so as to have nearly equal levels of fluency; and although it was not possible to have an equal number of each sex in *each* group, the total of 216 data sets were for an equal number of men and women.

The range of average number of "Yes" responses for the 18 sequences was from 980 to 1,678. In an analysis of variance the differences among the 18 sequences were significant ($F = 3.664$, 17 and 198 *df*, $p < .01$). Two conclusions are possible: (*a*) the groups were unequal in their disposition to judge the trigrams as meaningful and/or (*b*) the particular sequences of viewing altered this disposition to judge the item as meaningful.

In either case it seemed imprudent to ignore the effect of sequence and use a disproportionate number of *S*s from any one sequence. Therefore all sequences were given equal weight, and only the data for the *S*s appearing in the just described split-plot analysis of variance were used in the final determination of association values. Regrettably this meant using the data for only 216 *S*s rather than the 319 cases available. Though this may appear to be grossly inefficient, it seemed desirable in order to assure a reasonably good unbiased estimate of association value. The loss in precision was not very great: assuming an association value of 50%, the $\sigma_{\%}$ is 3.4 for $n = 216$, whereas it would be 2.8 for $N = 319$. Even the former is a considerable improvement over Glaze ($\sigma_{\%} = 12.9$, for $N = 15$).

Practice in making the required judgments also proved to have a significant effect upon the likelihood of regarding a trigram as meaningful. The mean percentages of trigrams judged meaningful as a function of the order of viewing film strips were: 54.1, 53.6, 53.5, 52.8, 53.6, 54.4, 56.1, 57.0, and 56.8, from first to last, respectively. Though the change was not great, it was significant ($F = 5.150$, 8 and 112 *df*, $p < .001$). This result suggests that *S*'s criteria shifted with practice; and since the terminal values are higher than the initial ones, it may also be that *S* became more imaginative and willing to judge an item as meaningful.

Because of the matching operations performed in constructing the sequence groups, it would be meaningless to test for a sex difference in association value for the matching group data. Probably a better test is to compare the initial, unselected sample of 319 cases. Women were slightly more inclined to judge a trigram as meaningful than were men, 1,402.7 vs. 1,358.4, respectively. The difference was insignificant, however ($t = 1.024$, 317 *df*, $p > .05$).

To illustrate the nature of the individual differences observed in the present study, the frequency distribution of *S*s judging various numbers of trigrams as meaningful (out of a possible 2,480) is shown in Figure 1. For simplicity the data were grouped in intervals of 50.

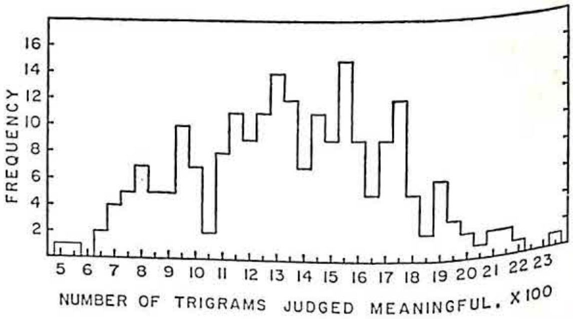


FIG. 1. A frequency distribution of the number of trigrams judged meaningful (maximum possible was 2,480) for the 216 *S*s (108 men and 108 women) whose data appear in the final tabulations. (Since there was no overall difference between the sexes, a single distribution is shown.)

The main results of the present study are shown in Table 1. The trigrams are arranged alphabetically within each level of association value found in the present study. All of these values are based on the responses of 108 men and 108 women.

TABLE 1

ASSOCIATION VALUES OF TRIGRAMS

Trigram	Archer	Glaze	Krueger	Trigram	Archer	Glaze	Krueger	Trigram	Archer	Glaze	Krueger
XYH	1	0	19	XYJ	2	0	21	XEJ	3	0	18
XAZ	2	13	39	XYQ	2	7	23	XIW	3	0	16
XIH	2	0	30	GYQ	3	0	34	XOV	3	13	36
XIJ	2	7	29	QUJ	3	0	22	XUF	3	13	23
XOJ	2	7	14	QYJ	3	0	18	XUJ[a]	3	13	22
XYB	2	0	36	XEF	3	0	23	XUQ	3	7	19
								XUY	3	0	29

[a] Significant sex difference.

TABLE 1—Continued

Trigram	Archer	Glaze	Krueger
XYF	3	0	22
XYV	3	0	30
YIJ	3	13	24
ZOJ	3	0	41
ZYJ	3	[b]	19
KYJ	4	7	38
QOJ	4	7	29
XIY	4	7	19
XOY	4	27	47
XUH	4	0	24
XUV	4	0	26
XYW	4	7	30
YEJ	4	27	54
YEQ	4	33	35
ZIJ	4	27	46
JYH	5	7	45
JYQ	5	0	41
MYV	5	0	43
QIJ	5	0	40
QYV	5	7	59
VYJ	5	0	24
XAH	5	7	27
XAJ	5	7	42
XEG	5	13	34
XEH	5	0	23
XIF	5	7	41
XIQ	5	7	39
XOZ	5		30
XUK	5	0	35
XUW	5	0	32
XUZ	5		34
XYC	5	20	32
XYD	5	7	37
XYS	5		42
YEV	5		36
ZUJ	5	13	30
GYX	6	27	52
JYK	6	13	54
QEJ	6	33	24
QIH	6	0	40
QYG	6	20	32
QYH	6		38
VUQ	6	0	25
WYJ	6	27	62
XEQ	6	0	42
XEV	6	27	28
XEY	6	13	35
XIB	6	33	36
XIK	6	33	55
XIZ	6	27	30
XOF	6	20	29
XOQ	6	20	53
XOS	6		58

Trigram	Archer	Glaze	Krueger
XUD	6	7	51
XUG	6	20	42
XUS	6	20	50
XYK	6	0	34
XYM	6		30
XYN	6	47	40
XYR	6		32
ZEJ	6	0	42
ZIH[a]	6	7	55
ZYX	6	20	56
BYJ	7	13	54
CYJ[a]	7	20	42
DYJ	7	40	51
GYJ	7		37
HYJ	7	13	61
KUJ	7	33	48
KYH	7	0	51
LYJ[a]	7	27	49
NIJ	7	20	55
NYJ	7	7	58
QYF	7	0	43
QYW	7	20	39
SYJ	7	13	46
VYF	7	0	64
XAD	7	0	42
XEW	7	20	48
XEZ	7	7	36
XOH	7	20	40
XOK	7	20	43
XOP	7	27	45
XUL	7	7	51
XUP	7		37
XYG	7	7	28
YIV	7		23
ZYF	7	13	49
ZYH	7	13	33
ZYW	7	0	31
CIJ	8	0	44
JYC	8	0	42
KIJ	8		51
QEF	8	13	33
QIY	8	13	41
QOH	8	20	57
TYJ	8	33	44
WUJ	8	7	39
WUQ	8	0	44
XAB	8	7	45
XAF	8	20	31
XAW[a]	8	13	57
XEK	8		40
XUB	8	7	43
XUT	8	20	55
ZAJ	8	7	51

Trigram	Archer	Glaze	Krueger
ZIY	8	13	42
ZUF	8	7	54
ZUY	8	7	36
BYW	9	—[c]	52
GEJ	9	20	40
GEX	9	0	38
KYF	9	7	46
KYG	9	7	56
PYV	9	7	63
QUV	9	27	52
QUX	9	33	43
QYB	9	13	50
TEJ	9	40	52
VIJ	9	40	43
VUJ	9	27	49
XEC	9	33	40
XID	9	13	50
XOC	9	20	45
XOG	9	27	39
XOL	9	27	47
XON	9	13	55
XOT	9	27	74
XUN	9	53	51
XUR	9	0	45
XYT	9	7	53
YAJ	9	93	
YAV	9	7	49
YIX	9	7	34
YOJ	9	20	23
YUF	9	0	49
YUJ	9	7	37
ZYQ	9	0	42
ZYS	9	0	70
FAJ	10	33	55
FYQ	10		51
GYC	10	13	—[c]
HIJ	10		74
KYV	10	0	42
LUJ	10	47	59
MUJ	10	33	78
PYB	10	7	33
QAZ	10	—[c]	53
QEX	10	13	55
QIW	10	7	44
RYW	10	47	58
TUJ	10	33	62
VOF	10	27	30
VUF	10	7	39
XIN	10	27	50
XIC	10	40	42
XAQ	10	7	31
XOM	10	33	41
XUC[d]	10	0	31

---

[b] Blank space indicates trigram was not evaluated.

[c] Item has two values.

[d] Significant change in meaning.

TABLE 1—Continued

Trigram	Archer	Glaze	Krueger	Trigram	Archer	Glaze	Krueger	Trigram	Archer	Glaze	Krueger
YIG	10	7	—[c]	DUJ	13	33	74	YUX	14	20	44
YOF	10	0	23	GIJ	13	47	54	ZEH	14	13	46
ZOQ	10	27	44	GOQ	13	0	54	ZUX	14	20	41
				JIH	13	20	44	ZYR	14	13	55
CYW	11	27	50	JIQ	13	7	60				
FEJ	11	47	66	JYT	13	27	77	BIW[a]	15	33	50
FEQ	11	27	46	JYX	13	—[c]	69	BUV	15	40	55
FOJ	11	20	62	KYB	13	20	57	CIW	15	20	57
GAQ	11	—[c]	61	MOJ	13	33	60	CUJ	15	20	57
GEQ	11	40	41	MYP	13	13	47	FYP	15	20	53
GIX	11	7	47	MYW	13	7	59	GIQ	15		44
GOJ	11	27	58	NUQ	13		63	KIH	15	13	63
GUJ	11		55	PYW	13		66	NEJ	15	20	55
GUQ	11	0	41	QEH	13	40	39	NUV	15	53	68
GYK	11	7	42	QEP	13	20	65	POJ	15		73
JYD	11		38	QUW	13	7	48	QOK	15	27	72
KUQ	11	13	63	TIJ	13	40	52	QUC	15	47	68
QAJ	11	27	37	VYH	13	40	29	RIW	15	40	51
VOJ	11	20	58	WIJ	13	40	61	VOQ	15	33	58
VYG	11	27	65	WOJ	13	7	45	VYQ	15	0	46
XEM	11	60	66	XED	13	33	53	VYT	15	20	65
XET	11	27	61	XEP	13	40	56	XEN	15	33	65
XIG	11	13	53	XIL	13		54	YIZ	15	27	61
XIM	11	27	41	XUM	13	0	60	ZEQ	15	0	35
XYL	11	13	55	YEB	13	27	30	ZOV	15	13	54
YEF	11	27	47	YEC	13	27	51	ZYK	15	13	45
YIQ	11	13	37	YIH	13	13	31				
ZOF	11	7	55	ZIW	13	0	31	BOJ	16	27	70
ZYC	11	13	56	ZOS	13	7	43	COJ	16	27	51
				ZUV	13	7		DIJ	16	40	64
CUQ	12	40	63	ZYB	13	7	46	DYB	16	27	
FYJ	12		59	ZYD	13	0	53	FYH	16	47	50
GOX	12	27	40	ZYV	13	0	43	GIW	16	0	67
JYZ	12	0	49					JYS	16	7	69
KOJ	12	7	50	FEP	14	20	45	KEJ	16	20	55
KYW	12	13	52	FYM	14		62	NYD	16		63
NOJ	12	13	63	HUJ	14	53	72	QEB	16	13	48
NYV	12	27	62	JEQ	14	7	48	QEC	16		48
QEV	12		44	JIY	14	27	74	RUV	16	47	50
QIF	12	0	44	LAJ	14	0	70	SUJ	16	53	61
QUH	12	—[c]	42	LYW[d]	14		65	VAW	16	47	51
SYW	12	47	56	QAH	14	27	33	VOH	16	13	39
TYV	12	33	43	QAV	14	40	68	VUH	16	7	38
TYW[d]	12	13	53	QIG[a]	14	47	55	VYR	16	47	60
VAJ	12	40	53	QOV	14	0	30	WUC	16	20	42
VEF	12	27	45	QUG	14	0	38	XAV	16	7	72
VUB	12	60	43	SIJ	14	0	53	YIB	16	13	39
VYB	12	7	53	VUP	14	33	44	YIF	16	7	56
WYB	12	20	58	VYW	14	7	47	YOX	16	7	49
XAL	12	7	53	WYV	14	7	66	YUQ	16	7	29
XER	12	60	55	XAN	14	53	54				
XIR	12	53	44	XAS	14	60	57	CEJ	17	7	38
XOB	12	13	38	XAT	14	0	49	FUV	17	33	49
XOD	12	40	50	XES	14	—[c]	65	GAC	17	53	72
XYP	12		41	XIS	14	40	38	GEC	17	7	42
ZUQ	12		39	YIW	14	7	38	GYW	17	7	67
				YOQ	14	7	41	JUQ	17	7	64
DEJ	13	20	36	YUB	14	13	46	KIF	17	27	47

TABLE 1—Continued

Trigram	Archer	Glaze	Krueger
LIJ	17	27	53
LYG	17	40	78
MEJ	17	13	63
MIV	17	13	51
PEJ	17	33	49
QAG	17		53
QAS	17	7	56
QEK	17	40	66
QEZ	17	40	75
QOB	17	0	43
QYX	17		55
RYX	17	13	57
TOJ	17	40	53
VEQ	17	13	57
VUG	17	73	64
VYZ	17	20	54
WUX	17	—[c]	58
XAK	17	40	49
XAY	17	20	59
XEB	17	7	22
YOB	17	47	62
ZEC	17	20	51
ZIQ	17	0	34
BYV	18	13	62
FYC	18	7	
FYD	18	33	78
GYH	18	20	33
HAJ[a]	18	40	64
HYF	18	40	
JUF	18	20	43
JYR	18	33	74
KYZ	18	—[c]	80
MYB	18	20	45
MYJ	18	7	56
NAJ	18	53	56
QOG	18	60	50
QYS[a]	18	27	—[c]
RUJ	18	93	78
RYV	18	40	80
TIW	18	27	65
VUX	18	7	54
VUY	18	27	60
VYD	18	87	59
XAC	18	40	
XIV	18	0	42
YOZ	18	7	51
CEQ[a]	19	33	51
DYW	19	13	68
FYK	19	20	—[c]
FYW	19	20	51
GAX	19	0	55
GIC	19	13	44
GUX	19	27	43
HIW	19		51
HYG	19	67	77
JOF	19	33	72
JYW	19	7	51
KIW	19	33	48
KUY	19	13	58
LIW	19	20	69
NYW	19	27	63
PYJ	19		59
QOL	19		56
QOS	19	7	50
QOX	19	20	43
QOZ	19	53	71
QUF[d]	19		60
QUK	19	20	54
SOJ	19	33	75
VOB	19	27	48
VUK	19	0	44
VUW	19	27	52
WYQ	19	33	75
XAG	19	33	—[c]
XOR	19	27	59
XOW	19	47	46
YAF	19	47	—[c]
YEX	19	33	60
ZEV	19	20	56
ZYL	19	13	69
FIQ	20	47	76
JIC	20	0	51
NYZ	20	53	48
PYM	20	47	74
QED	20	20	72
QOF	20	40	66
VIH	20	20	69
WEF	20	20	67
WEQ	20	20	53
ZIF	20	0	60
ZUH	20	7	43
BIH[d]	21	47	47
BUW[a]	21	47	67
HYB	21	47	71
JEV	21	47	56
JIK	21	13	45
PIJ	21	47	73
QEM	21	7	52
QOC	21	0	46
QOP	21	27	43
VIF	21	100	61
WUB	21	0	54
XEL	21	53	73
YIL	21	0	57
ZIX	21	20	69
ZYT	21	0	65
CAQ	22	40	59
DAJ	22	33	44
FIH	22	40	70
FIW	22	27	64
GIH	22	20	41
GUH	22	20	57
HOJ	22		58
KEH	22	20	62
KYQ	22	27	49
MEF	22	0	44
MIW	22	40	56
NIY	22	40	59
NYH[d]	22	33	74
PUV	22	40	61
QAB	22	40	60
QAX	22	27	78
QET	22	47	72
QOW	22	7	49
RIH	22	27	57
RYJ	22	13	74
RYQ	22	20	58
SUV	22	—[c]	71
TUV	22	13	68
TYB	22	53	66
VAB	22	47	46
VAF[d]	22	0	57
VEJ	22	13	57
VUD	22	27	47
WIH	22	47	59
WOY	22	33	65
XAP	22	20	45
YOV	22	0	39
ZAF	22	27	48
ZEF	22	20	65
ZUW	22	20	27
ZYN	22		73
CIH[d]	23	40	38
CUG[a]	23	40	61
CYH	23		49
GYZ	23	20	52
HIF	23	20	61
JEX	23	0	58
KEQ	23		64
KIY	23	27	56
LOJ	23	27	74
MYZ	23	13	77
QEG	23	27	54
QON	23	20	54
QUL[a]	23	27	73
QYC	23	73	79
QYM	23		42
TYG	23		72
VYX	23	7	76
WAJ	23	20	57
WYX	23	20	63
YAB	23	13	50
ZAX	23	13	60
ZIK	23	7	61
ZOB	23	13	43
BEJ	24	20	57
BIQ	24	40	76

TABLE 1—Continued

Trigram	Archer	Glaze	Krueger
BUH	24	27	66
BYQ[a,d]	24	27	64
CUH	24	13	68
GUW	24	20	53
HYW	24	20	51
JOM	24	33	65
KIV	24	—[c]	59
KYR[a]	24	33	73
MUB	24	47	62
PYD	24	20	53
QAP	24	0	57
QIV	24	67	71
QUB[d]	24	7	48
QUY[d]	24	33	52
RUQ	24	13	57
SEJ	24	33	72
YUG[d]	24	47	59
YUT	24	93	58
ZAV	24	33	68
ZOK	24	13	70
ZYM	24	—[c]	64
BIJ	25	53	58
BYP	25	33	43
CEH	25	13	45
CYB	25	40	60
CYZ	25	47	70
DAX	25	0	48
FIY	25	67	67
FOQ	25	13	64
FUP	25	20	38
FYV	25	7	74
HUQ	25	20	73
JEH	25	13	61
JYV	25	0	46
KAQ[a]	25	13	57
KEB	25	7	62
KOQ	25	33	55
KUW	25	40	57
MYF	25	7	57
NUY	25	47	63
NYQ	25	13	65
PEF	25	40	40
PYF	25	60	79
QIR	25	67	85
QOR[a]	25	47	67
QUN	25	—[c]	55
QYD	25	27	73
QYZ	25	33	68
RIJ	25	27	70
RUW	25	20	72
TOV	25	0	62
TYH	25	27	77
VUC	25	33	58
WOQ	25	33	60
XAR[a]	25	13	49
XIT	25		65
XYZ	25	47	79
YAD	25	47	43
YOD	25	33	63
ZAH	25	40	39
ZAQ	25	27	69
ZAS	25	13	55
ZEG	25	7	38
ZUB	25	7	60
ZUL	25	53	59
ZUN	25	40	77
BOQ	26	53	70
GIY	26	20	63
HYV	26	47	64
MAF	26	27	57
MUX	26	27	73
MYQ[a]	26	27	72
NIH	26	33	73
NIZ	26	27	63
NYB	26	20	78
PYH[d]	26	33	55
QAF	26	47	70
QES	26	87	85
QYR	26	27	74
SIW	26	20	71
VUR	26	33	61
XIP	26	—[c]	71
YED	26	27	59
YEK[d]	26	7	41
YUV[a]	26	0	37
ZID[d]	26	20	64
ZOC	26	27	59
ZOH	26	7	60
ZOY	26	—[c]	44
ZUG	26	47	—[c]
BAV	27	73	68
FEK	27	60	64
FOV	27	53	77
GEH	27		55
GYS	27	73	68
JUV	27	73	74
LEJ	27	27	75
MEQ	27	0	63
MIB	27	7	58
NAX	27	13	71
QOM	27	13	60
QUP	27	20	58
RUX	27	33	65
TOQ[a]	27	53	68
VAH	27	40	64
VEK	27	40	61
VIW	27	47	75
VOZ[a]	27	47	69
VYS	27	47	65
WUG	27	13	51
YAZ	27	7	79
YUD	27	0	39
ZEY[d]	27	20	66
ZOL	27	27	72
ZUK	27	0	63
CEF	28	0	64
DYH	28	27	41
DYT	28	47	78
LEQ	28	20	64
MIH[a]	28	47	70
NAF	28	33	62
PUJ	28	47	73
QUD	28	47	62
QYK[d]	28	40	88
VAQ	28	53	76
VUS	28	40	58
YIC	28	20	42
ZAL	28	27	65
ZEW	28	40	67
ZIB	28	20	57
ZOT	28	7	64
ZUC	28	27	61
CUY	29	53	71
GEF	29	13	71
JEG	29	20	66
JUH	29	7	71
JUY	29	47	83
KEZ	29	7	51
KUG	29	40	57
KUV	29	40	76
NIV	29	33	74
PIH	29		63
QIB	29	47	64
QUS	29	20	69
RYH	29	33	77
TUZ	29	33	68
VEH[a]	29	—[c]	61
VUT	29	47	60
WUK	29	33	73
WYD	29	53	85
WYH	29	20	67
YOH	29	13	63
ZOX	29	27	68
ZUD	29	27	58
CAJ	30	40	72
DOJ	30	53	73
FAQ	30	33	77
GIK	30	13	46
GYD	30		67
KAJ	30	13	59
KAX	30	27	62
KIG	30	20	65
KUH	30	20	64
LUY	30	20	72
NUX[a]	30	7	79
NYR	30	33	77
PUY	30	40	65
QYP	30	40	66

TABLE 1—Continued

Trigram	Archer	Glaze	Krueger
RUY	30	20	56
SYZ	30	33	75
VEZ	30	60	63
VUM	30	20	66
WUH	30	0	57
YAX	30	53	62
YIK	30		32
ZAB	30	33	53
ZAW	30	13	79
BEX	31	—[c]	52
BOF	31	7	52
BYH	31	33	67
CYQ	31	27	49
CYX[a]	31	27	63
DUY	31	53	75
FUH	31	7	76
FUY	31	33	71
FYS	31	53	78
GAK	31	27	65
GEV	31	53	75
GIZ	31	47	67
HYN	31		62
KAH	31	33	70
KEF	31	20	57
LUH	31	40	72
MEV	31	27	39
MIY	31	33	78
MYH[a]	31	40	68
NYM	31	67	70
PIW	31	7	73
POB	31	47	55
PYQ	31		75
QEN	31	27	55
QUR	31	40	67
SAJ[a]	31	20	72
SYH	31	27	66
TAJ	31	47	70
TEF	31	20	52
VEB	31	47	58
VEY	31	73	70
VIQ	31	—[c]	70
VOP	31	27	69
WUV	31	33	53
YEG[a]	31	80	86
YIM	31	0	50
YOL	31	27	63
ZAY	31	33	70
ZED	31	27	72
ZET	31	53	54
ZIR	31	13	64
ZOD	31	33	64
ZUR	31	20	66
BUJ	32	27	82
CEB	32	27	75
CEG	32	53	69
CYF	32	33	60
CYV[d]	32	47	64
DIW	32	20	72
FIJ	32	80	74
FIK	32	40	65
GEK	32	27	40
HUV	32	80	79
JIW	32	13	73
KIQ	32	53	82
KOH	32	40	68
KUC	32	87	65
LYH	32	33	67
MEH	32	40	63
NAQ	32	27	64
NOQ	32	47	76
QEW	32	7	67
QEY	32	20	70
QYL	32	27	79
RIY	32	27	72
RUH	32	47	86
TAZ	32	67	60
TEH	32	33	70
TUH	32	20	75
TYF	32	20	79
VYL	32	73	68
YAG	32	27	53
ZIC	32	20	58
CAX[a]	33	40	73
DYF	33	47	78
FEG	33	47	69
GAJ	33	33	83
HYX	33	27	66
LYB	33	53	86
MEB	33	40	72
NYF	33	47	67
NYG	33	60	72
PYZ	33	27	64
QOD	33	13	75
SYQ	33	20	63
TIH	33	40	69
TUQ	33	33	76
TUY	33	47	78
TYS	33	40	69
VYP	33	27	65
WIB	33	33	58
WYK	33	53	75
WYS[a d]	33	53	76
ZEM	33	53	61
ZIS	33	13	76
ZYG	33	0	57
BYM	34		78
CUV	34	73	79
DIH	34	33	73
DIY	34	60	79
FUQ	34	27	81
GYF	34	27	68
JOZ	34	33	72
KEC	34	33	59
LIH	34	27	71
MIQ[a]	34	53	74
MUH	34	33	70
MUP	34	40	75
NIQ	34	27	79
QAR	34	47	89
QOT	34	53	82
TAH	34	20	62
VIY	34		60
VUZ	34	20	68
WOX	34		65
YEM	34	0	64
YIN	34	13	—[c]
YUW	34	20	63
YUZ	34	0	73
ZER	34	20	62
ZUP	34	7	82
ZUS	34	20	55
BAJ	35	53	73
BYS	35	53	73
BYX	35	40	59
CIQ	35	33	77
COH	35	27	82
DEQ	35	40	77
GAH	35	0	62
JID	35	0	64
JUB	35	67	77
JUW	35	47	81
KUX	35	33	72
MOQ[a]	35	40	91
NUC	35		77
NUK	35	20	74
PEH	35	27	59
PIB	35	33	43
QIX	35	53	74
QOY	35	13	71
QUM	35		47
QYN	35	—[c]	73
SIH	35	20	78
SUQ[d]	35	27	87
SUW[a]	35	67	73
TEB	35	53	68
TEY	35	33	81
TIV	35	20	51
VEC	35	0	67
YIS	35	60	79
YIT	35	47	71
YUN[d]	35	20	67
ZEX	35	0	63
ZOG	35	13	65
BUP	36	47	61
CAZ	36	53	64
COQ	36	47	77
CUX	36	13	59
DYQ	36	13	68

TABLE 1—Continued

Trigram	Archer	Glaze	Krueger
GED	36	13	64
HOQ	36		87
JAF	36		73
JEC	36	13	68
JIZ[a]	36	27	68
MEP	36	47	61
NIW	36	47	61
NYS[d]	36	33	71
QER	36		82
QUZ[a]	36	20	77
REH	36	40	73
ROJ	36	27	73
RUZ	36	27	61
SEH	36	33	73
TEQ	36		82
VOR	36	33	62
WUY	36	47	75
ZES	36	40	71
ZOP	36	20	70
BIY	37	33	77
BOV	37	73	75
FEC	37	47	70
GYB	37	27	79
GYT	37	40	75
HYZ	37	13	75
JUC	37	13	69
JUX	37	27	63
JUZ	37	20	75
KYP	37	33	81
LYQ	37		81
MIJ	37	47	69
PUX	37	40	68
QAT	37	53	76
QAW	37	60	64
QIM	37	7	58
TEG	37	27	74
VAK	37	20	84
VED	37	33	50
ZAK	37	53	77
ZIL	37	0	71
BAZ	38	40	71
BUQ	38	20	71
CEV	38	33	73
CUW	38	40	61
CYG	38		61
DAQ	38	7	54
DYV[a]	38	67	88
FEH	38	0	55
FUW	38	47	75
HEQ	38	27	78
LOQ	38	53	92
MAQ[a]	38	47	85
MUQ	38	33	77
NIR	38	47	79
NOF	38	40	70
NUH[a]	38	53	68
PAZ	38	60	74
REZ	38	33	77
RIQ	38	53	—[c]
RYS[a]	38	53	85
SIY	38	60	79
SOZ[a]	38	47	69
SYV	38		82
TEZ	38	47	65
TUW	38	27	75
TYX	38	33	84
VYC[a, d]	38	0	82
VYN[d]	38	40	83
WEC	38	60	61
WEM	38	53	69
WEX[a]	38	53	86
WIQ	38	47	76
WOG	38	7	67
WUM	38	60	84
XAM	38	87	88
YOG	38	40	55
YOP	38	0	56
YOS	38	40	73
YOT	38	13	66
ZAD	38	47	72
ZEN	38	47	58
ZUT	38	13	58
CIZ[a]	39	33	66
CYM	39	60	75
DOY	39	80	77
GIS	39	67	83
GOC	39	33	62
GYV	39	33	74
HUY	39	40	70
HYQ	39	7	77
JEK	39	40	79
JOW	39	73	85
JUP	39	73	74
KAZ	39	20	67
KEX	39	13	57
LUW	39	47	73
NEZ	39	67	71
PIY	39	13	66
QUT[a]	39	40	81
REJ	39		70
TUP	39	67	76
TYQ	39	20	67
TYZ[a]	39	13	71
VAD	39	7	76
VYK	39	67	71
WAB	39	67	77
WEZ	39	7	72
WYC	39	60	79
ZIV[a]	39	7	60
CAH	40		74
CEK	40	40	66
CIY	40	60	64
FUB	40	13	50
GOK	40	20	72
GUB	40	27	69
GYR	40	60	84
JUK[a]	40	60	76
JYL[a, d]	40	47	91
KYC	40	40	66
LOZ[a]	40	53	79
MOF	40	40	83
MYD	40	—[c]	80
PAF	40	13	41
PEX	40	40	75
QAN	40	67	77
RYL	40	60	82
TEV	40	13	43
TIY	40	60	86
TOZ[a]	40	40	69
VIB[a]	40	13	63
VUN[d]	40	33	72
WYZ	40	53	90
YUS[d]	40	20	85
BEQ[a]	41	67	68
BYF	41	33	84
BYZ	41	73	79
DIQ	41	27	80
FOZ	41	60	74
GEP	41	33	63
GOH[a]	41	47	74
HEZ	41	33	71
HIG	41	87	73
HUW	41	13	69
JAT[a]	41	20	61
JIR	41	53	65
KAG[a]	41	27	71
LYF	41	80	93
NEM	41	60	85
NIS	41	60	60
POQ	41	33	72
QAY	41	40	85
QYT	41	47	82
SEF	41	60	66
SUY	41	33	69
SYF[a]	41	40	83
TIQ	41	53	87
VAX	41	80	83
WEH	41	47	76
BEW	42	47	64
FEX	42	47	81
FIM[d]	42	73	62
FIP	42	53	75
GEB	42	20	64
GEZ	42	47	71
GOW	42	47	80
GUD	42	40	89
GUV[d]	42		86
LUQ	42	53	87

TABLE 1—Continued

Trigram	Archer	Glaze	Krueger
MIP	42	33	58
NEQ	42	40	81
PUH[a]	42	40	72
PUQ	42	40	83
QID	42	20	78
RYZ[a]	42	40	85
VAZ	42	33	84
VOK	42	53	73
WEJ[a]	42	40	80
WIY[a]	42	27	76
FOH[a]	43	33	56
FUJ	43	40	85
FYZ	43	60	81
GOY	43		77
HIY	43	60	79
PUW	43	33	72
QEL	43	60	78
RIX	43	13	63
SYB	43		82
VAY	43	53	75
WIV	43	73	81
YAN	43	40	64
ZAN	43	40	65
ZON	43	33	74
BEH	44	20	63
BEZ	44	60	77
DEH	44	27	71
DOQ	44	33	87
DUT	44	80	86
DYX	44	33	82
DYZ	44	60	88
FAP	44	53	60
HIB	44	73	77
JAH	44		75
JOR[d]	44	80	80
JYB	44	27	65
KIB	44	53	79
NID	44	33	60
NUJ[d]	44	27	80
PEQ	44	47	78
PIV	44	53	75
QAM	44	20	69
QIS	44		—[c]
RAJ	44	87	90
RUK	44	47	76
RYC[a]	44	67	63
RYF	44	67	80
RYK	44	27	70
SEQ	44	67	86
WIX[a]	44	27	87
WYR	44	60	78
YIR	44	40	66
YOC	44	27	62
DEG[a]	45	67	71
GOZ	45	60	72
HAQ[a]	45	47	86
HEG	45	20	81
JOH	45	60	80
JOX	45		69
MOG	45	40	84
MOY	45		73
MUY	45	40	70
MYG	45	67	66
PEM	45	33	68
POV	45	53	73
QIC	45	73	83
TEP[a]	45	87	86
WOB	45	53	71
WYG	45	27	83
WYM	45	60	78
YUH	45	27	75
CYK	46	33	76
DYS	46	67	76
HUZ	46	47	92
KEV	46	20	46
KYS[a]	46	60	88
LIG	46	73	89
LIX	46	—[c]	90
LIY[a]	46	73	87
LYS[a]	46	53	92
NEP	46	73	86
PEV	46		66
TOH	46	33	74
TYD	46	60	83
VEM	46	67	81
VOX	46	60	81
YAT	46	47	79
ZEP	46	53	91
BYK	47	67	73
CIB	47	27	64
FUX[a]	47	33	87
GOM	47	67	87
GUC	47	13	46
HIQ	47		88
JYF	47	40	76
LYX	47	73	91
MYX[d]	47	53	86
QAC	47	73	90
QIZ	47	87	
RAX	47	67	90
RUC	47	33	65
RYG	47	53	86
SIQ	47	53	89
SUG	47	87	89
TOF	47	53	89
TUD[d]	47	47	68
TUL	47	87	82
VYM	47	47	89
WUP	47	13	72
YAQ[a][d]	47		52
ZIN	47	13	69
CIF[a]	48	67	69
CUK	48	60	78
CYS[a][d]	48	33	89
DUQ	48	67	85
DYR	48	73	86
GYL	48	60	89
HEJ	48	27	76
HIV[d]	48		83
KAC[a]	48	73	67
KIZ	48	20	78
MUW	48	13	77
MYC[a]	48	33	72
MYK[a]	48		82
NOY	48	47	79
PAJ	48	53	65
POY	48	53	79
QAD	48	40	—[c]
QAL	48	73	79
REW	48	73	74
SEB	48	13	51
VEP	48	20	65
VUL	48	60	93
WYF	48	40	88
BIV	49	40	60
BYG	49	60	83
KYL[a]	49	40	93
LEB	49	20	62
MUZ	49	73	87
NEF	49	47	72
NUS	49	60	77
QIP[a]	49	60	76
SOQ	49	47	86
WYL[a][d]	49	—[c]	88
BEM	50	60	65
CYR	50	53	81
DIB[a]	50	40	77
DUH[a]	50	27	60
FAH	50	47	70
FOW	50	80	85
GUK	50	7	50
HYL	50	60	84
JIS	50	47	74
JYN	50	87	95
KES	50	60	83
KYN[a]	50	60	85
KYX	50		78
LAH[a]	50	47	77
LOH	50	67	80
LUF	50	67	94
MAB	50	80	81
MOX	50	20	76
NUZ[d]	50	53	68
POZ[a]	50	47	81
PYR	50	80	87
QIK	50	67	91
QIN	50	33	75

TABLE 1—Continued

Trigrams	Archer	Glaze	Kreuger
ROH	50	67	82
SAH	50	60	81
SOH[a]	50	53	87
SYX	50	67	80
VID	50	40	57
VOD	50	47	68
WOH	50	40	83
BAQ[a]	51		83
BYC	51	73	85
CEP	51		76
DOH	51		75
HAX	51	47	89
JOP	51	67	80
NAZ	51	80	80
PEZ[a]	51	33	67
SUF	51		95
SUX	51	40	86
TAQ	51	67	91
VIZ	51		92
YOM	51	7	63
YUC	51	27	53
BYR	52	47	87
DAK	52	47	72
FAW[d]	52	67	76
FID	52	73	94
FOT	52	80	86
GEY	52	47	85
GIF	52	67	86
HEB[a]	52	47	80
HEF	52	47	87
JEY	52	33	65
JIV[d]	52	27	54
JOQ	52	27	88
KUL	52		84
LEH	52	47	77
LEK[d]	52	87	84
LEZ	52	33	84
MEZ	52	20	84
MIF	52	7	60
NOH[a]	52	53	73
NUW	52	27	89
NYK[d]	52	53	80
POH[a]	52	20	75
ROQ	52	87	91
SAZ	52	60	84
SYG	52	60	68
TAY	52	60	89
VOS	52	67	76
ZAR	52	67	79
ZEK[a]	52	73	60
BOH[a]	53	60	78
CYL	53		87
DEZ	53	67	67
DUW[a]	53	40	77
FEV[d]	53	93	88
JYG	53	47	77
LER	53	53	86
LUP	53		84
LYV[a]	53	60	87
MEK	53	67	88
NIF	53	60	88
NIM	53	67	74
POF	53	67	58
POG	53	—[c]	66
PYT	53	73	90
PYX	53	47	88
VIT	53	67	88
WUL[a]	53	80	77
ZIM	53	47	64
ZIT	53	33	79
CAG	54	73	85
FYL[d]	54	67	87
FYX	54	60	83
GAW	54	27	67
KEW[a]	54	73	81
KOS	54		95
KUN	54		80
LUT	54	67	89
MYN[a]	54	60	89
NUP	54	73	73
PID	54	33	65
PUZ	54	93	79
QAK	54	73	96
RAQ[a]	54	73	91
RYN	54	80	87
SOV[d]	54	93	89
WEV[d]	54	100	78
WOS[a]	54	20	74
YEZ	54	53	95
ZYP	54	33	90
FAZ	55	47	82
FOC[a]	55	73	73
HOZ	55	33	83
HYT	55		88
JIT	55	67	84
JOD	55	33	73
JUM	55	87	90
MUV	55	60	80
MYS[a]	55	93	94
NAS	55	80	91
PIQ	55	73	89
SUH	55	40	87
SYK	55	67	91
TUS	55	80	84
VES[d]	55	100	90
WUD	55	80	80
WYT	55	53	87
ZAT	55	13	83
ZEL	55	33	74
ZUM	55	0	74
BUX	56	87	93
CIM	56	73	82
CYP[a]	56	53	76
DEV	56	80	92
FYB	56	60	92
JAV	56	87	91
KOV[a]	56	60	71
KYM	56	27	70
LYM[d]	56	67	89
PAQ	56	27	88
PEB	56	87	89
PYC[d]	56		90
WEG	56	40	86
WYP	56	53	81
BYN	57	67	86
CES[a]	57	47	78
DYP	57	53	—[c]
FAV	57	93	92
HYC[a]	57		75
HYR[a]	57	47	81
JEZ	57	20	82
LYC[a]	57	80	89
MOH	57	40	84
MYL[a]	57	67	87
NAL	57	40	73
PIF	57	73	92
PIM	57	60	75
TYC	57	—[c]	91
WOV	57	73	84
ZAC	57	27	77
BEP	58	13	64
BYT[d]	58	53	82
DYL[a]	58	53	88
JEP	58	47	77
JIX	58		81
KET	58	87	93
LAQ	58	73	97
LOY	58	93	87
LUD	58	40	85
MAV	58	60	74
MIZ	58		88
MYR[a]	58	80	92
NYP	58	40	87
PYK	58	53	83
QIL	58	100	89
RYB	58	67	82
SEK	58	80	87
SYM	58	73	94
SYT[a]	58	60	90
TAS	58	80	87
WEK[a]	58	60	72
YID	58	40	84
BAF	59	73	89
BOC	59	87	82
BYL[a]	59	7	85

TABLE 1—Continued

Trigrams	Archer	Glaze	Kreuger
CEW	59	60	79
CEY	59	40	78
CIX[a]	59	60	67
DOF	59	73	77
DUV[a]	59	67	86
DYC	59		92
FAX	59	53	91
FYG	59	67	84
GAF[a]	59	27	78
HIX[a]	59	40	92
KOZ[a]	59	80	82
LIR	59	60	92
MOZ	59		84
NYT[a]	59	80	93
SYC	59	73	80
TAW[a]	59	40	76
TIB	59	73	78
TIG	59	100	82
CYT[a]	60	80	92
DOX	60	73	85
FYT	60	73	90
GAZ	60		94
HYK[a]	60	60	86
KOX	60		85
LYT[a]	60	73	91
NEH	60	27	83
NER	60	73	93
PYS	60		78
SUL	60	87	95
VAP	60	53	89
WEP	60	53	80
WOC	60		59
ZAM	60	53	71
BAX	61	67	90
BIX	61	27	58
DYG	61	73	82
GAN	61	80	80
GEW	61	40	73
GUZ	61	80	88
HYS	61	60	88
JAL	61	53	85
LYR	61	60	94
NEX	61	73	87
PUM	61	73	89
RUP	61	67	84
RYT	61	80	91
SAQ	61	60	89
TYK[a]	61	40	92
VIG	61	27	84
YAR	61	93	—[c]
ZEB	61	53	73
BAP	62	67	72
BIK	62	67	90
BYD[a][d]	62		94
CEZ	62	47	81
DAC	62	40	89
FIS	62	80	87
FOM	62	80	75
FUG	62	67	89
FUT	62	93	86
JUR[a]	62	87	92
KUF	62	53	87
LYK[a][d]	62	60	97
MOK[a]	62	67	87
MYT	62	67	95
NAD	62	60	70
PAH	62	40	85
TEM[a]	62	73	89
TUR	62	93	87
VOM	62	53	86
YAL	62	67	85
YUR[a]	62	53	86
BEF	63	73	82
BIC	63	87	83
BOD	63	80	93
BOZ	63	53	77
CEM	63	60	81
FYR	63	93	96
GID	63	—[c]	77
GOS	63	80	88
GUF	63	27	79
GUP	63	67	70
HIN	63	93	87
HOV[d]	63	87	90
JOL	63	93	93
JOV	63	67	93
LEC	63	93	79
LUS[d]	63	87	93
LUZ[d]	63	60	77
LYP[a][d]	63	80	95
LYZ	63	60	87
NUD	63	67	73
NYX	63	40	86
REQ	63	67	86
SEV	63	93	93
SIZ	63	73	95
TYR	63	73	92
VEN	63	67	93
WAZ	63	47	77
WOL	63	87	90
YAS	63	60	99
ZOM	63	13	67
CEX	64	13	69
DAH[a]	64	33	74
GOF	64	67	83
HAB	64	67	94
KAV[d]	64	73	80
KEM	64	47	80
KYT	64	80	91
LIF	64	100	98
NOM	64	93	90
PYL[a]	64	87	94
RIK	64	53	83
RIS[d]	64	87	85
SAB	64	87	90
SIF[a]	64	40	90
TIX	64	73	89
TOX	64	87	93
VIL	64	93	88
WAH	64	47	83
WAQ	64	60	79
WOF	64	73	76
YOK	64	93	85
FOL	65	93	95
FON	65	93	92
HET	65	80	92
HUD	65	87	93
JEB[a][d]	65	33	59
KAF	65	93	87
KEL	65	87	88
KIR	65	73	78
KOC	65	47	73
KYD[d]	65	73	93
LEP	65	87	96
LOD	65	80	94
LOM[d]	65	73	81
LOX	65	80	87
MEY	65		91
NAR	65	100	84
NYC	65		89
PAB	65	60	60
PYG	65	80	92
PYN[a][d]	65	60	93
SYP	65	100	90
VIR	65	93	89
VIX	65	40	93
VOG	65	40	85
WID	65	80	90
WOK	65	67	86
WUS[a]	65	47	80
BEK[a]	66	47	94
FIV	66	93	92
GES	66	87	89
JAD	66	73	85
LUN	66	93	93
LYD	66	93	92
NUG	66	93	91
PIZ	66		59
QIT	66	87	96
RIZ[a]	66	73	85
RYD	66	53	92
TOB	66	73	87
VOC	66	27	78
CAK[a]	67	87	89
CIK[a]	67	33	84
DAP	67	80	89

TABLE 1—Continued

Trigrams	Archer	Glaze	Kreuger
DAW	67	80	89
DUR	67	87	92
DYN	67		94
FUM[a]	67	80	90
HUK	67	60	81
KAS	67	87	83
LEF[a]	67	87	95
MEC	67	7	85
MEX[a]	67	100	96
MUN	67	47	94
NUR[a]	67	53	86
NYL[a]	67	73	90
PEY	67	87	92
RAL[d]	67	100	80
ROF	67	53	84
TIR[a, d]	67	93	89
TYN	67	67	94
VER	67	87	94
WIK	67	40	90
WOM	67	100	92
WUF	67	60	86
YUK	67	40	59
BIM	68		86
CUL	68	87	96
DEY	68	87	88
DIR	68	—[c]	89
FEY	68	60	70
FIC	68	87	91
FOS	68	80	96
FYN	68	80	92
GUR	68	73	78
JAQ[a]	68	67	87
JOS	68	80	86
JYP	68	67	94
LAZ	68	93	97
PES	68	100	93
RET	68	87	95
SAR	68	100	90
SEG	68	67	88
SUZ[a]	68	67	65
TYL[a]	68	80	94
VAM	68	73	71
VEW	68	80	81
VOY	68	73	88
WOZ[d]	68	27	74
BOK	69	87	89
CET[a]	69		81
DET	69	93	97
DYK[a]	69	87	92
FES	69	100	86
FOD	69	80	89
HOD	69	87	93
HUC	69	7	80
JYM	69	87	96
KOM	69	80	89
LUB	69	47	88
NEG	69	87	90
NUF[a]	69	93	93
RIL	69	80	83
RIV	69	87	95
SAF	69	87	86
SOF	69	100	96
SYL	69	87	96
SYR	69	73	96
TOK	69	87	95
WIR	69	87	87
WUR	69	67	82
WUT	69	20	70
DYM	70	73	95
HEV	70	73	88
HUS	70	73	97
KEP	70	87	88
KUB	70	53	79
LOF	70	87	98
RYP[a]	70	80	97
SOT	70		
SUT	70	100	92
TAV	70	60	71
VON	70	47	79
WOD	70	87	89
YER	70	67	84
ZAP	70	33	81
BIP[a]	71	20	57
CER	71	87	86
CIV	71	67	87
FAS	71	100	93
FEZ	71		
HES	71	27	87
KOF	71	80	89
LIQ	71	100	95
NEB	71	80	86
TAF	71	67	97
BIS	72	93	81
CIL	72		93
CYN[a]	72	87	97
DAS	72	60	86
DOK	72	73	96
HYD	72		98
LAV	72	80	92
MAZ	72	80	97
NAC[a]	72	47	90
NAV	72	100	91
NES	72	80	95
NOP	72	60	88
REL	72	73	95
TID	72		93
VOT	72	67	93
WEY[a]	72	73	89
CED[a]	73	60	83
COV	73	93	95
CUN[d]	73	60	78
DUS	73	93	98
GYN	73	87	83
JUN	73	80	97
KOD[a]	73		80
LUM[d]	73	80	84
NAM	73	87	85
REK	73	87	90
RUD	73	100	98
SIV	73	73	87
YOW	73	20	85
CAY	74	67	85
COS	74	93	94
COZ	74	93	96
DAF	74	53	96
DAL	74	80	93
DUP	74	53	85
DUX	74	60	89
FEN[a]	74	73	95
HEK	74		90
HOF	74	87	86
HOX[a]	74	73	91
KOL	74	80	91
MAJ	74	80	89
MUL	74	73	86
NAK	74		89
PAG	74		85
PIR[a]	74	80	91
SIB	74	60	74
TEK	74	40	88
TYM[a]	74	67	96
VIK	74	100	95
WAP[a]	74	33	91
BEY	75		
CAS	75	67	97
CEN	75	100	91
CIP	75	67	74
CYD[a]	75	53	88
DES	75	60	90
DOB	75	100	82
FOP	75		
HIZ	75	60	84
HOK	75	73	91
HUP	75	80	97
HUX	75	60	88
KOR	75	47	89
LAN[a]	75	93	97
LIB[a]	75	93	91
MOS	75	87	95
NOK	75	—[c]	98
PEL	75	80	86
TIZ[a]	75	47	68
VAG	75	73	94
VIS	75	100	91
BIF	76	80	80

TABLE 1—Continued

Trigrams	Archer	Glaze	Kreuger
DAZ[a]	76		96
FAK	76	93	94
JUS	76	100	97
KER	76	53	89
KOG	76	40	74
LAR	76	93	88
MAH	76	73	93
MER	76	87	91
MUR[a]	76	93	92
NUB	76	13	73
PAV	76	93	86
RAB	76	93	93
RES	76	87	94
ROP	76	100	96
RUL[a]	76	93	83
TEW[a]	76	27	80
TIF[a]	76	80	89
VAS	76	93	96
YAW	76		
BUK	77	80	89
CAV	77	80	96
FAC	77	87	92
FET	77	60	94
GAV	77	67	87
KUS	77	60	94
LEV	77	93	94
RYM	77	93	93
WIF	77	73	93
YEH	77	80	85
BIZ	78	100	95
COK[d]	78	73	93
DEP	78	100	95
DEX	78	73	89
FOK	78	80	83
GOL[a][d]	78	87	93
KAL	78	33	93
KOW	78	60	96
KOY	78	33	89
KUZ[a]	78		86
LIM	78	100	98
MIR	78	93	94
MOV	78	87	93
NOL	78	47	83
POS	78	93	94
REN[a]	78	80	86
TEC[d]	78	80	86
YON	78	80	91
YOR	78	47	91
ZOR[d]	78	20	74
DIF	79	100	97
FAL	79	93	97
FUK	79		87
HOY	79		88
JAX	79	47	87
JES	79	93	95
KIP	79	80	91
NOC	79	87	—[c]
NUL	79	93	89
PAX[d]	79	80	83
REY[a]	79	80	95
RIF	79	87	94
SOG	79	73	86
SUK	79	60	95
TER	79	100	95
VEG	79	87	84
BEC[a]	80	100	92
CIS	80		97
DER	80	87	91
GAT[a]	80		92
GIB	80	87	90
GIM	80	67	96
JAS	80	93	96
JOK	80	93	96
KAW	80	53	86
KOB	80	53	83
KON	80	53	86
KUD	80	53	82
LEX	80		90
MIK	80	80	95
MOD	80	100	96
NEV	80	87	95
RAD[d]	80	100	95
RAV[d]	80	100	96
ROV	80	100	94
ROZ[a]	80	67	89
TES	80	100	98
TIS	80	47	88
TUK	80	87	95
TYP[a]	80	93	97
YEW	80		93
BIR	81	87	91
COF	81	93	98
FOY	81	33	77
FUC[a]	81		94
GER	81	87	92
GIR	81	87	87
HAF	81	73	89
HOS	81	93	90
HYP	81	87	99
JAK	81	73	96
JUL	81	—[c]	93
KUR	81	67	93
LEM	81	80	87
LEY	81	67	91
MIC[d]	81	53	95
MUS	81	93	99
NEK[a]	81	93	96
NIB	81		80
NIG[a]	81	93	97
NOX	81	67	81
NOZ[a]	81	80	94
RAF	81	80	96
SEM[d]	81	100	90
SIM	81	67	95
TOS[a]	81	93	95
VAC	81	100	97
WOR	81	87	93
WOT	81		94
WUN	81	60	96
ZIG	81	60	82
ZOW	81	7	83
CIR	82	93	86
COL	82	100	97
DEK[a]	82	93	90
DIZ	82	80	97
DOP	82	93	97
FAM	82	100	99
GON	82	93	97
GUL	82	100	95
HAK	82	67	93
JER	82	93	93
JIB	82	60	84
LOK	82	80	97
MUC	82	80	91
PUD	82	93	95
ROK	82	73	94
ROM	82	100	95
VIN	82	100	93
VOL	82	93	97
BER[a]	83	100	92
BOL	83	100	96
BOT[a]	83	87	91
DAG	83	67	90
DIV	83	93	95
DUF	83	47	90
FEB	83	73	81
FOB[a]	83		
GAR	83		99
GOR	83	80	93
HAR	83	73	93
KAB[a]	83	47	87
KUM	83	—[c]	89
LAT	83	80	93
LIS	83	100	95
LOR	83	80	93
NEY	83	53	91
NOG	83	67	83
POC	83		92
PON	83	87	94
PUC	83		95
PUK	83	47	83
RAS[a]	83	93	95
SYD[a]	83	60	96
VAR	83	93	85
WAF	83	73	90
WAM	83	67	90
WIZ[a]	83	87	94

TABLE 1—Continued

Trigrams	Archer	Glaze	Kreuger
WYN	83		98
BAW	84	60	95
DIS	84	100	97
DIT	84	80	85
FER	84	80	93
HIR	84	93	91
HUL	84	80	98
HUR	84	100	99
KED[a,d]	84	47	92
MOC	84	87	95
NOS[a]	84	93	98
NOV	84	100	88
PEK	84	87	97
POK	84	93	95
REM	84	93	100
SEP	84	100	99
SIK	84	80	96
TIK	84	80	97
TOC	84	53	88
WAL	84		97
WER	84	73	92
WES	84	80	91
BES	85	93	96
GEN	85	87	98
HAN	85	87	97
KAD	85	73	94
KAM	85	47	92
MUK	85	60	91
POM[a]	85	80	92
RAK	85	87	98
TAL	85	100	93
TOR	85	87	94
WIC	85	80	96
WIM[a]	85	87	91
CIG	86	87	96
DAV	86	67	91
DOS	86	93	89
FUD	86	80	87
GYP	86		97
HYM	86		92
JOC	86	73	96
KOT	86	73	97
KUT	86	60	92
LUC	86		100
LUK	86	93	95
MOT	86	80	95
REB	86	60	91
RUF[a]	86	100	99
SAV	86	93	98
SED	86	87	99
SOR[a]	86		97
WIP	86	87	95
WUZ	86		91
YAC	86	93	67
YUP[d]	86	27	62
BUF	87	73	91
CUF	87	93	97
GIP	87	60	95
HOL	87	100	98
KUP	87	67	90
LYN[a]	87	87	94
MAK	87	100	97
MOL	87	93	95
NIK	87	60	97
ROS	87	93	95
SER	87	100	96
SEZ	87	67	95
SYN[a]	87	100	97
TOL	87	80	98
TUC[a]	87	80	98
VOW	87		
WAK	87	100	88
BUC	88	80	93
COG	88		
COR	88	87	96
DAR	88	100	95
DEF	88	93	91
HEC	88		96
HIK	88	87	97
HOC	88		93
JEF	88	73	94
JEN[a]	88	87	93
JIP	88		91
LAS	88	87	99
LOP	88		
LUR[a]	88	80	96
LUV	88	73	96
NIT	88	93	99
RUS	88	100	96
SEC	88	93	93
SEN	88	80	97
SEY[a]	88	67	92
SOK[a]	88	73	99
SOM[a]	88	80	99
TAK[d]	88	100	99
VIP	88	53	77
WAN	88		
YEP	88	80	93
YUL	88	53	86
ZAG	88	60	72
BAK	89		94
CIN	89	100	96
DOM	89	93	99
DOW	89	80	94
DUK[a]	89	87	98
GEL	89	80	97
GIV	89	100	99
HED	89	67	92
JAC[a]	89	87	98
JED	89	53	72
JIN	89	100	99
KIC	89	60	91
KOP[a]	89	67	98
LIC	89	93	97
LON	89	100	98
MES	89	60	97
MUF	89	93	97
NAH	89	13	85
NEC[a]	89	93	98
PEC[a]	89	73	96
POL	89	93	99
SUD	89	93	96
TUF	89	87	93
WAV[a]	89	87	94
YIP	89		89
BUL	90	100	97
CAF	90		95
CID[a]	90	80	87
DAT	90	73	94
DIK	90	80	95
DOV	90	100	98
DUC	90	100	96
FEM	90	93	97
FIZ	90		93
FUS	90	93	97
GOP	90	53	95
HAC	90	73	89
HEX	90	53	86
HOB	90	100	87
JEM	90	80	94
JUT	90		
LAC	90		97
LAK[a]	90	93	99
LOC[a]	90	93	100
PER	90		
PIS	90		92
POR[a]	90		95
PUL	90	93	100
ROG[a]	90	87	94
SEL	90	100	88
SUC	90	73	98
VEX	90		
WEN	90	80	94
BAC	91	—[c]	93
BAL	91	100	97
BUR	91	100	92
CAW	91		
CEL	91	80	95
CUR	91		
DUL	91	100	98
HAW	91	80	98
HOM	91	100	99
HOR	91	87	95
HUF	91	87	93
JIF	91	67	96
LIK	91	100	96
MAS	91	93	99

TABLE 1—Continued

Trigrams	Archer	Glaze	Kreuger
NEL	91	80	95
PIK	91	80	98
REG	91	100	94
RIC	91	93	88
ROL	91	67	98
ROX	91	73	95
SUR	91	100	98
YAM	91		
BOM	92	80	98
BOR	92	93	94
COM	92		97
CUM	92	93	94
DIX	92		97
DOR	92	87	98
FUL	92	80	97
HAP	92		99
KAR	92	53	95
KIL	92	93	99
LEW	92		98
LOB	92	93	97
NIC	92		97
NOR	92		
PAS	92	100	99
PED[a]	92	100	92
PUF[a]	92	100	100
RIN	92	87	88
SAN[a]	92	80	98
SIG	92	87	94
SIL	92	93	95
YEL	92	93	99
BAS	93	100	97
BOS	93		95
CAM	93		
CIT	93		94
CUS	93	100	98
CUZ	93	73	91
FEL	93	100	99
FIL	93	80	94
HAV	93	100	97
KIS	93	87	99
LAF	93	87	100
LIN	93	87	95
LIV	93	100	98
LOS	93	100	97
MOW	93		
NUM	93	67	90
PAC	93	100	97
PAK	93	87	98
PIL	93	100	99
RAC	93	100	95
REC	93	80	91
REV	93	100	96
VAL	93		96
WOP	93		92
YAH	93	27	88
BAH	94		94
BEV	94	100	90
BIL	94	93	93
COX	94		
DEM	94	73	98
DIC	94	100	91
DIL	94	80	94
DOL	94		100
FAG	94		
FUZ	94	93	96
HAZ	94	100	95
HEW	94		
HIL	94	100	99
JAB	94		
KAN	94	87	96
KAP	94		96
KID	94		
LOV	94	100	100
MAL	94	93	98
MIN	94	93	99
MOR	94		97
NAW	94	47	91
PIX	94	67	88
PUR	94	93	96
RAZ	94	87	99
REF	94	93	93
REP	94		98
RIT	94		98
ROC	94	87	97
SAK	94	93	98
SOL	94	93	100
SOP	94		
TUN	94	73	98
VIM	94		
WAT	94	100	95
WEL	94		95
WIS	94	100	94
CUD	95		
DIN	95		
DUN	95		
GAM	95		96
GIL	95	87	94
GOV	95	93	95
HEL	95		99
JAZ	95	93	97
JIL	95	87	98
JOG	95		
JUD	95		86
KIX	95	47	80
LAM	95	87	99
LEN	95	87	92
MAG	95	87	96
MEG	95	67	90
MEW	95		
MIG[a]	95	33	77
MIL	95	93	96
POX	95		97
SIC	95	93	93
VAT	95		
YAP	95	40	89
BOG	96		
BON	96		97
CAL	96		
DEC	96	100	94
DUB	96		
DUM	96		99
FIR	96		
HUN	96		
LIT[a]	96		99
LUX	96		99
MUT	96		
NAB	96		
NAT	96		99
SOY	96	80	87
SUP	96		100
VEL	96	87	92
BAM	97		93
CAD	97		
COB	97		
COY	97		
DEB	97	87	99
DEL	97	80	97
FAB	97	93	92
GAD	97		98
GIT	97		97
GOB	97	93	97
GUS	97		100
HEY	97		
HON	97	100	98
HUG	97		
HUM	97		
JEL	97	100	98
MAW	97		
MEL	97	67	95
MIS	97	80	99
MON	97	100	98
NIX	97		
PAM	97	80	90
PAW	97		
PEW	97		
POD	97		
PUS	97	100	
SOC	97	100	98
TAC	97	100	98
TAD	97		96
TAX	97		
TIL	97	80	100
TOG	97	73	97
TOW	97		
TUX	97	80	96
VIC	97	100	97
WAS	97		
WIL	97	100	96

TABLE 1—Continued

Trigrams	Archer	Glaze	Kreuger
WON	97		
YAK	97	53	
YEN	97		90
YET	97		
YUM	97	33	89
BOP	98	53	69
COT	98		
DOC	98		100
DOZ	98	100	94
FAY	98		
GAB	98	93	100
GYM	98	100	97
HEM	98		
HEN	98		
HIS	98		
HUB	98		
JAW	98		
JAY	98		
JIG	98		
JOT	98		
KAT	98		100
KIM	98	60	89
LAX	98		
LED	98		
LES	98	87	97
LUG	98		
NAY	98		
NIL	98	87	97
NIP	98		
NOB	98		91
PUN	98		
REX	98		
SAC	98		99
SAL	98		97
SAX	98		99
SID	98	87	97
SOX	98		99
SUB	98	100	91
TAM	98	100	100
TOD	98		
TON	98		
TUM	98	73	97
BAD	99		
BAG	99		
BEL	99	100	97
BEN	99		
BID	99		
BIG	99		
BOW	99		
BOX	99		
BUG	99		
BUS	99		
BUT	99		
BUY	99		
BUZ	99	93	96
CAR	99		
COD	99		
CUB	99		
DAB	99		
DAN	99		
DEW	99		
DIM	99		
DUZ	99	47	80
FAN	99		
FAT	99		
FED	99		
FEW	99		
FIB	99		
FIG	99		
FIN	99		
FOG	99		
FUR	99		
GAL	99		98
GAP	99		100
GAY	99		
GOD	99		
GOT	99		
GUM	99		
GUT	99		
GUY	99		
HAG	99		
HAL	99		100
HEP	99		87
HIC	99		96
HID	99		
JAG	99	67	97
JAN	99	100	99
JAP	99		
JAR	99		
JON	99	93	96
KEG	99		
KEY	99		
KIN	99		
LAD	99		
LAY	99		
LEG	99		
LID	99		
LIP	99		
LOW	99		
MAC	99	80	100
MAN	99		
MAR	99		
MAX	99	80	98
MID	99	87	99
MIX	99		
MOB	99		
MOP	99		
MUG	99		
NAG	99		
NED	99		
NEW	99		
NOW	99		
PAN	99		
PAR	99		
PAT	99		
PAY	99		
PEG	99		
PEN	99		
PIC	99	93	96
PIG	99		
PIT	99		
POT	99		
PUG	99		96
PUT	99		
RAG	99		
RAH	99	60	100
RAN	99		
RAP	99		
RAW	99		
RED	99		
RID	99		
RIG	99		
RIM	99		
ROD	99		
RON	99	80	85
ROW	99		
RUB	99		
RUN	99		
RUT	99		
SAT	99		
SAY	99		
SET	99		
SEW	99		
SIN	99		
SIP	99		
SIT	99		
SIX	99		
SOD	99		
SOW	99		
SUN	99		
TAN	99		
TEL	99	93	99
TEN	99		
TIC	99	100	98
TIM	99		
TOM	99		
TOY	99		
TUG	99		
WAC	99	67	89
WAD	99		
WAG	99		
WAY	99		
YES	99		
BAN	100		
BAR	100		
BAT	100		
BAY	100		
BED	100		
BEG	100		
BET	100		
BIN	100		

TABLE 1—Continued

Trigrams	Archer	Glaze	Kreuger
BIT	100		
BOY	100		
BUD	100		
BUM	100		
BUN	100		
CAB	100		
CAN	100		
CAP	100		
CAT	100		
CON	100		97
COP	100		
COW	100		
CUP	100		
CUT	100		
DAM	100		
DAY	100		
DEN	100		
DIG	100		
DIP	100		
DOG	100		
DON	100		
DOT	100		
DUG	100		
FAD	100		
FAR	100		
FIT	100		
FIX	100		
FOR	100		
FOX	100		
FUN	100		
GAS	100		
GEM	100		
GET	100		
GIN	100		
GUN	100		
HAD	100		
HAM	100		
HAS	100		
HAT	100		
HAY	100		
HER	100		
HIM	100		
HIP	100		
HIT	100		
HOG	100		
HOP	100		
HOT	100		
HOW	100		
HUT	100		
JAM	100		
JET	100		
JEW	100		
JIM	100		
JOB	100		
JOY	100		
JUG	100		
KAY	100	60	97
KEN	100	100	99
KIT	100		
LAB	100		100
LAG	100		
LAP	100		
LAW	100		
LET	100		
LIZ	100	80	94
LOG	100		
LOT	100		
MAD	100		
MAP	100		
MAT	100		
MAY	100		
MED	100	93	98
MEN	100		
MET	100		
MIT	100		
MUD	100		
NAP	100		
NET	100		
NOD	100		
NOT	100		
NUT	100		
PAD	100		
PAL	100		
PET	100		
PIN	100		
POW	100	60	95
PUB	100	93	85
RAM	100		
RAT	100		
RAY	100		
RIB	100		
RIP	100		
ROB	100		
ROT	100		
ROY	100		
RUG	100		
RUM	100		
SAD	100		
SAG	100		
SAM	100		
SAP	100		
SAW	100		
SEX	100		
SIR	100		
SOB	100		
SON	100		
SUM	100		
TAB	100		93
TAG	100		
TAP	100		
TAR	100		
TED	100		
TEX	100	100	97
TIN	100		
TIP	100		
TOP	100		
TUB	100		
VAN	100		
VET	100	93	99
WAR	100		
WAX	100		
WEB	100		
WED	100		
WET	100		
WIG	100		
WIN	100		
WIT	100		
ZIP	100		

[a] Significant sex difference.
[b] Blank space indicates trigram was not evaluated.
[c] Item has two values.
[d] Significant change in meaning.

In the "Glaze" and "Krueger" columns a blank space (Footnote b) indicates that Glaze or Krueger did not evaluate the item. Both authors deliberately excluded most three-letter words. Footnote c in either of these columns indicates that two different values for the item were reported by each author. Glaze had 29 duplications (and most of them were not eliminated in Hilgard's chapter), and Krueger had 14 duplications.

In the "Trigram" column Footnote a means that there may be a significant difference between the sexes for that particular trigram. A chi square (corrected for discontinuity) for the 2 × 2 contingency table of sex × meaning was computed for all 2,480 items. If the chi square exceeded 3.841 (the value needed at the .05 level for 1 *df*), the trigram is annotated with Footnote a. There are 226 such items. Theoretically we might expect 124, i.e., .05 × 2,480 chi squares to be significant at the .05 level by chance. Therefore, it is unlikely that there is a true sex difference for all 226 trigrams listed in Tables 2 and 3.

TABLE 2

192 TRIGRAMS WHICH APPEARED TO BE MORE MEANINGFUL TO WOMEN

Trigram	Association Value			Chi Square	Trigram	Association Value			Chi Square
	Overall	Women	Men			Overall	Women	Men	
BAQ	51	59	44	4.744	HYR	57	65	49	4.834
BEC	80	86	73	4.823	JAC	89	94	84	4.866
BEK	66	76	56	9.065	JAQ	68	78	57	9.321
BEQ	41	50	32	6.192	JEN	88	95	81	9.838
BER	83	89	78	4.033	JUK	40	48	32	4.927
BOH	53	61	44	5.368	JUR	62	72	51	9.470
BOT	83	90	77	5.633	JYL	40	51	30	9.315
BYD	62	75	49	14.331	KAB	83	89	78	4.033
BYL	59	69	49	7.644	KAC	48	56	39	6.013
BYQ	24	33	15	9.144	KAG	41	49	32	5.542
CAK	67	76	57	7.521	KAQ	25	31	19	4.173
CAX	33	40	26	4.112	KED	84	91	78	5.899
CED	73	80	66	4.570	KEW	54	62	45	5.381
CEQ	19	25	13	4.335	KOD	73	81	66	5.303
CES	57	65	49	4.834	KOP	89	94	84	4.866
CET	69	76	61	4.829	KOV	56	64	49	4.238
CID	90	96	84	7.596	KOZ	59	67	51	4.892
CIK	67	74	60	4.112	KUZ	78	89	67	14.170
CIX	59	67	52	4.315	KYL	49	56	42	4.168
CIZ	39	46	31	4.383	KYN	50	58	41	6.001
CYD	75	87	64	14.402	KYR	24	31	18	4.280
CYJ	7	11	3	4.585	KYS	46	57	34	10.741
CYN	72	81	63	7.402	LAH	50	59	41	6.685
CYP	56	66	47	6.799	LAK	90	95	85	5.275
CYS	48	56	41	4.172	LAN	75	82	68	5.556
CYT	60	69	51	6.975	LEF	67	74	60	4.112
CYX	31	38	23	4.909	LIB	75	82	69	4.901
DAH	64	73	56	6.539	LIT	96	100	93	6.361
DAZ	76	82	69	4.280	LIY	46	55	37	6.042
DEK	82	90	75	7.185	LOC	90	94	85	4.099
DIB	50	57	42	4.741	LOZ	40	47	32	4.347
DUH	50	63	37	13.500	LUR	88	94	82	5.291
DUK	89	94	84	4.866	LYC	57	65	50	4.260
DUV	59	70	48	10.144	LYJ	7	12	3	5.468
DUW	53	61	44	5.368	LYK	62	70	54	5.681
DYK	69	76	62	4.241	LYN	87	94	79	10.197
DYL	58	67	49	6.152	LYP	63	70	56	4.467
DYV	38	44	31	3.872	LYS	46	54	39	4.190
FEN	74	81	66	6.101	LYT	60	68	53	4.347
FOH	43	50	35	4.260	LYV	53	61	44	5.368
FUM	67	76	57	7.521	MAQ	38	47	29	7.096
GOH	41	50	32	6.192	MIH	28	35	20	5.192
GOL	78	84	72	3.916	MIQ	34	41	27	4.056
HAQ	45	54	36	6.063	MOK	62	71	52	7.827
HIX	59	67	52	4.315	MOQ	35	46	24	10.739
HOX	74	81	68	4.074	MUR	76	84	69	6.571
HYC	57	66	49	5.472	MYC	48	60	36	11.590
HYK	60	68	52	4.927	MYH	31	38	23	4.909

TABLE 2—Continued

Trigram	Association Value			Chi Square
	Overall	Women	Men	
MYK	48	58	38	8.178
MYL	57	66	49	5.472
MYN	54	62	46	4.774
MYQ	26	32	19	4.074
MYR	58	67	50	5.505
MYS	55	62	47	4.203
NAC	72	80	65	5.192
NEC	89	94	83	5.672
NEK	81	89	74	6.903
NOH	52	60	44	4.751
NOS	84	92	77	7.854
NOZ	81	88	73	6.650
NUF	69	77	60	6.203
NUH	38	46	29	6.400
NUR	67	75	58	6.021
NYL	67	80	55	14.183
NYT	59	67	51	4.892
PEC	89	95	83	7.007
PED	92	99	85	12.514
PEZ	51	58	44	4.168
PIR	74	81	67	4.671
POH	52	59	44	4.172
POM	85	92	78	7.010
POR	90	94	85	4.099
POZ	50	59	40	7.408
PUF	92	96	88	4.086
PUH	42	50	33	5.505
PYL	64	72	56	5.799
PYN	65	72	58	4.003
QOR	25	36	15	11.806
QUL	23	30	17	4.398
QUT	39	49	30	7.759
QUZ	36	43	29	3.956
QYS	18	25	11	6.133
RAQ	54	61	46	4.190
RAS	83	89	78	4.033
REN	78	84	71	4.527
REY	79	89	69	12.181
RIZ	66	75	57	6.704
ROG	90	94	85	4.099
ROZ	80	86	74	4.181
RUF	86	92	80	5.424
RUL	76	82	69	4.280
RYC	44	53	34	6.799
RYP	70	78	62	5.634
RYS	38	48	27	9.560
RYZ	42	50	33	5.505
SAJ	31	40	22	7.010
SAN	92	96	88	4.086
SEY	88	93	82	4.233
SOH	50	61	40	8.964
SOK	88	93	82	4.233
SOM	88	94	81	7.389
SOR	86	92	80	5.424
SOZ	38	49	27	10.399
SUW	35	42	28	4.003
SUZ	68	78	58	8.518
SYD	83	90	76	6.392
SYN	87	93	81	5.736
SYT	58	66	50	4.861
TAW	59	68	50	6.192
TEM	62	69	54	5.009
TEP	45	53	38	4.203
TEW	76	82	69	4.280
TIF	76	82	69	4.280
TIR	67	74	59	4.688
TIZ	75	83	68	6.401
TOQ	27	34	20	4.570
TOS	81	90	72	9.754
TOZ	40	49	31	6.975
TUC	87	92	81	3.983
TYK	61	75	47	16.383
TYL	68	77	59	6.900
TYM	74	81	67	5.424
TYP	80	89	71	9.408
TYZ	39	48	31	6.285
VOZ	27	35	19	5.970
VYC	38	46	29	6.400
WAV	89	94	83	5.672
WEJ	42	50	34	4.861
WEK	58	66	51	4.286
WEX	38	45	31	4.423
WEY	72	81	63	7.402
WIM	85	94	76	11.589
WIX	44	51	36	4.238
WIY	42	49	34	4.286
WIZ	83	89	78	4.033
WOS	54	62	46	4.774
WUL	53	61	45	4.761
WUS	65	77	54	11.765
WYL	49	57	41	5.354
WYS	33	41	25	5.371
XAR	25	31	19	4.173
YAQ	47	55	40	4.180
YUR	62	70	53	6.340
YUV	26	32	19	4.074

TABLE 3

34 TRIGRAMS WHICH APPEARED TO BE MORE MEANINGFUL TO MEN

Trigram	Association Value			Chi Square	Trigram	Association Value			Chi Square
	Overall	Women	Men			Overall	Women	Men	
BIP	71	63	79	5.737	MEX	67	58	75	6.021
BIW	15	9	20	4.439	MIG	95	91	99	6.130
BUW	21	15	27	4.042	NIG	81	73	88	6.650
CIF	48	38	57	7.423	NUX	30	23	37	4.313
CUG	23	16	30	5.174	QIG	14	7	20	6.542
DEG	45	33	56	10.778	QIP	49	42	56	4.168
FOB	83	78	89	4.033	SIF	64	55	73	7.244
FOC	55	44	66	9.881	SYF	41	33	49	4.892
FUC	81	74	88	5.900	VEH	29	20	37	6.538
FUX	47	38	56	6.025	VIB	40	32	47	4.347
GAF	59	45	72	14.982	WAP	74	67	81	4.671
GAT	80	73	86	4.823	XAW	8	4	13	4.909
HAJ	18	11	25	6.133	XUJ	3	0	6	5.315
HEB	52	43	61	6.694	YEG	31	23	38	4.909
JAT	41	33	49	4.892	ZEK	52	44	59	4.172
JEB	65	58	72	4.003	ZIH	6	2	9	4.324
JIZ	36	28	44	5.799	ZIV	39	31	47	4.966

For 192 of the items for which a sex difference was indicated, the difference was in favor of women. As may be seen, the trigrams have been listed alphabetically. Table 2, which lists the trigrams which appeared to be more meaningful to women, shows the overall association values, values for the two sexes, and the chi squares (all are significant at or beyond the .05 level with 1 *df*). Table 3 shows comparable values for the 34 trigrams which appeared to be more meaningful for men.

Measures of reliability. One measure of the reliability of judgment was obtained from the test-retest correlation of the number of "Yes" responses made by *S* to the first film strip and to the identical strip which was repeated after *S* saw all of the others. For 16 of the groups this measure was based on a 300-item "test"; for two of the groups there were only 80 items in the "test." The Pearson product-moment correlation which describes this consistency of judgment over a 48-hr. period was .882 (214 *df*, $p < .01$).

Another measure of reliability was obtained by analyzing the responses for each individual trigram. Since two different sequences started with the same film strip, we have eight groups of 24 *S*s each judging 300 different trigrams twice (and one group of 24 *S*s making two judgings of the 80 trigrams of the short film strip). A fourfold point coefficient or tetrachoric correlation based on but 24 cases and many with zero cell entries is too unstable to be worthwhile. On the other hand a measure of the change in meaningfulness of a particular item could be evidence of reliability which would be stable since it would be less affected by zero cell entries. Such a measure of reliability was computed: a chi square test for a difference in number of "Yes" responses on the two "testing sessions." A chi square test for correlated values was used; and, if Footnote d appears in the "Trigram" column of Table 1, it indicates a significant *change* in meaningfulness over a 48-hr. period. Most of the changes were increases, i.e., more "Yes" responses on the second testing than on the first.

Either notation (Footnotes a or d) in the "Trigram" column is a caution signal to the potential user since it indicates either a sex difference or a lack of reliability.

Comparability of the three association values. One immediate question is: "Was the task worthwhile?" One answer to the question is the three overall correlations among the three studies, but this is a somewhat deceptive answer. When an *E* wants to construct a list of trigrams, he rarely samples from the entire possible population. The usual practice is to select items with a predetermined level of meaningfulness. Accordingly, a better answer to the question of the worthwhileness of this re-evaluation is to be found in the correlations of the three sets of values in the more restricted ranges from which an *E* might construct lists, e.g., high, medium, or low association values. In the present study there are 510 trigrams with an association value between 1% and 25% (none was found with 0%) for which Glaze and Krueger also report values. A Pearson product-moment correlation for these 510 items was computed between the association values found in the present study and each of the other two. For purposes of comparison, the Glaze and and Krueger values were also correlated. Comparable correlations were computed for each of the other quartile intervals. These values, along with related means and standard deviations, are shown in Table 4.

TABLE 4

CORRELATIONS AND MEANS FOR EACH OF FOUR RESTRICTED RANGES AND THE ENTIRE RANGE OF ASSOCIATION VALUES OF THE THREE SCALES

Range	N	r_{AG}	r_{AK}	r_{GK}	Means		
					Archer	Glaze	Krueger
1- 25%	510	.341	.590	.435	14.6	21.6	51.6
26- 50%	561	.354	.406	.489	37.4	39.5	71.9
51- 75%	440	.311	.341	.522	62.9	66.7	85.7
76-100%	422	.163	.286	.559	87.3	82.1	93.6
1-100%	1933[a]	.788	.850	.790	48.1	50.3	74.5

[a] After the elimination of Glaze's 29 two-valued items and Krueger's 14 two-valued items, there were 1,933 items for which all three studies reported a single value.

As may be seen from Table 4, although the overall correlations are fairly high, .788 to .850, the correlations for the restricted ranges are rather low, from .162 to .590, indicating that Glaze's and Krueger's values may not provide the best description of the meaningfulness of trigrams.

SUMMARY

The association values for 2,480 CVC trigrams were determined. Two measures of reliability of these values were also obtained. The association values obtained in the present study were correlated with those previously reported by Glaze and Krueger. The overall correlation was fairly high, but for four restricted, and more useful, ranges the correlations were rather low. It was concluded that the present scale offers a better estimate of the meaningfulness of trigrams than was previously available.

REFERENCES

GLAZE, J. A., The association value of nonsense syllables. *J. genet. Psychol.*, 1928, **35**, 255–269.

HILGARD, E. R., Methods and procedures in the study of learning. In S. S. Stevens (Ed.), *Handbook of experimental psychology*. New York: Wiley, 1951. Pp. 517–567.

KRUEGER, W. C. F. The relative difficulty of nonsense syllables. *J. exp. Psychol.*, 1934, **17**, 145–153.

(Received January 4, 1960)

Vol. 74, No. 11 Whole No. 498, 1960

Psychological Monographs: General and Applied

THE INFORMATION AVAILABLE IN BRIEF VISUAL PRESENTATIONS[1]

GEORGE SPERLING[2]

Harvard University

How much can be seen in a single brief exposure? This is an important problem because our normal mode of seeing greatly resembles a sequence of brief exposures. Erdmann and Dodge (1898) showed that in reading, for example, the eye assimilates information only in the brief pauses between its quick saccadic movements. The problem of what can be seen in one brief exposure, however, remains unsolved. The difficulty is that the simple expedient of instructing the observer of a single brief exposure to report what he has just seen is inadequate. When complex stimuli consisting of a number of letters are tachistoscopically presented, observers enigmatically insist that they have seen more than they can remember afterwards, that is, report afterwards.[3] The apparently simple question: "What did you see?" requires the observer to report both what he remembers and what he has forgotten.

The statement that *more is seen than can be remembered* implies two things. First, it implies a memory limit, that is, a limit on the (memory) report. Such a limit on the number of items which can be given in the report following any brief stimulation has, in fact, been generally observed; it is called the span of attention, apprehension, or immediate-memory (cf. Miller, 1956b). Second, *to see more than is remembered* implies that more information is available during, and perhaps for a short time after, the stimulus than can be reported. The considerations about available information are quite similar, whether the information is available for an hour (as it is in a book that is borrowed for an hour), or whether the information is available for only a fraction of a second (as in a stimulus which is exposed for only a fraction of a second). In either case it is quite probable that for a limited period of time more information will be available than can be reported. It is also true that initially, in both examples, the information is available to vision.

In order to circumvent the memory limitation in determining the information that becomes available following a brief exposure, it is obvious that the observer must not be required to give a report which exceeds his memory span. If the number of letters in the stimulus exceeds his memory span, then he cannot give a whole report of all the letters. Therefore, the observer must be required to give only a partial report of the stimulus contents. Partial reporting of available information is, of course, just what is required by ordinary schoolroom examinations and by other methods of sampling available information.

An examiner can determine, even in a short test, approximately how much the student knows. The length of the test is not so important as that the student not be told the test questions too far in advance. Similarly, an observer may be "tested" on what he has seen in a brief exposure of a complex visual stimulus. Such a test requires only a partial report. The specific instruction which indicates which part of the stimulus is to be reported is then given only after termination of the stimulus. On each trial the instruction, which calls for a specified part of the stimulus, is randomly chosen from a set of possible instructions which cover the whole stimulus. By repeating the interrogation (sampling) procedure many times, many different random samples can be obtained of an observer's performance on each of the various parts of the stimulus. The data obtained thereby make feasible the estimate of the total information that was available to the observer from which to draw his report on the average trial.

---

[1] This paper is a condensation of a doctoral thesis (Sperling, 1959). For further details, especially on methodology, and for individual data, the reader is referred to the original thesis. It is a pleasure to acknowledge my gratitude to George A. Miller and Roger N. Shepard whose support made this research possible and to E. B. Newman, J. Schwartzbaum and S. S. Stevens for their many helpful suggestions. Thanks are also due to Jerome S. Bruner for the use of his laboratory and his tachistoscope during his absence in the summer of 1957. This research was carried out under Contract AF 33(038)–14343 between Harvard University and the Operational Applications Laboratory, Air Force Cambridge Research Center, Air Research Development Command.

[2] Now at Bell Telephone Laboratories, Murray Hill, New Jersey.

[3] Some representative examples are: Bridgin (1933), Cattell (1883), Chapman (1930), Dallenbach (1920), Erdmann and Dodge (1898), Glanville and Dallenbach (1929), Külpe (1904), Schumann (1922), Wagner (1918), Whipple (1914), Wilcocks (1925), Woodworth (1938).

The time at which the instruction is given determines the time at which available information is sampled. By suitable coding, the instruction may be given at any time: before, during, or after the stimulus presentation. Not only the available information immediately following the termination of the stimulus, but a continuous function relating the amount of information available to the time of instruction may be obtained by such a procedure.

Many studies have been conducted in which observers were required to give partial reports, that is, to report only on one aspect or one location of the stimulus. In prior experiments, however, the instructions were often not randomly chosen, and the set of possible instructions did not systematically cover the stimulus. The notions of testing or sampling were not applied.[4] It is not surprising, therefore, that estimates have not been made of the total information available to the observer following a brief exposure of a complex stimulus. Furthermore, instructions have generally not been coded in such a way as to make it possible to control the precise time at which they were presented. Consequently, the temporal course of available information could not have been quantitatively studied. In the absence of precise data, experimenters have all too frequently assumed that the time for which information is available to the observer corresponds exactly to the physical stimulus duration. Wundt (1899) understood this problem and convincingly argued that, for extremely short stimulus durations, the assumption that stimulus duration corresponded to the duration for which stimulus information was available was blatantly false, but he made no measurements of available information.

The following experiments were conducted to study quantitatively the information that becomes available to an observer following a brief exposure. Lettered stimuli were chosen because these contain a relatively large amount of information per item and because these are the kind of stimuli that have been used by most previous investigators. The first two experiments are essentially control experiments; they attempt to confirm that immediate-memory for letters is independent of the parameters of stimulation, that it is an individual characteristic. In the third experiment the number of letters available immediately after the extinction of the stimulus is determined by means of the sampling (partial report) procedure described above. The fourth experiment explores decay of available information with time. The fifth experiment examines some exposure parameters. In the sixth experiment a technique which fails to demonstrate a large amount of available information is investigated. The seventh experiment deals with the role of the historically important variable: order of report.

---

[4] The experiments referred to are (cf. Sperling, 1959): Külpe (1904), Wilcocks (1925), Chapman (1932), Long, Henneman, and Reid (1953), Long and Lee (1953a), Long and Lee (1953b), Long, Reid, and Garvey (1954), Lawrence and Coles (1954), Adams (1955), Lawrence and Laberge (1956), Broadbent (1957a).

General Method

Apparatus. The experiments utilized a Gerbrands tachistoscope.[5] This is a two-field, mirror tachistoscope (Dodge, 1907b), with a mechanical timer. Viewing is binocular, at a distance of about 24 inches. Throughout the experiment, a dimly illuminated fixation field was always present.

The light source in the Gerbrands tachistoscope is a 4-watt fluorescent (daylight) bulb. Two such lamps operated in parallel light each field. The operation of the lamps is controlled by the microswitches, the steady-state light output of the lamp being directly proportional to the current. However, the phosphors used in coating the lamp continue to emit light for some time after the cessation of the current. This afterglow in the lamp follows an exponential decay function consisting of two parts: the first, a blue component, which accounts for about 40% of the energy, decays with a time constant which is a small fraction of a millisecond; the decay constant of the second, yellow, component was about 15 msec. in the lamp tested. Fig. 1 illustrates a 50-msec. light impulse on a linear intensity scale. The exposure time of 50 msec. was used in all experiments unless exposure time was itself a parameter. Preliminary experiments indicated that, with the presentations used, exposure duration was an unimportant parameter. Fifty msec. was sufficiently short so that eye movements during the exposure were rare, and it could conveniently be set with accuracy.

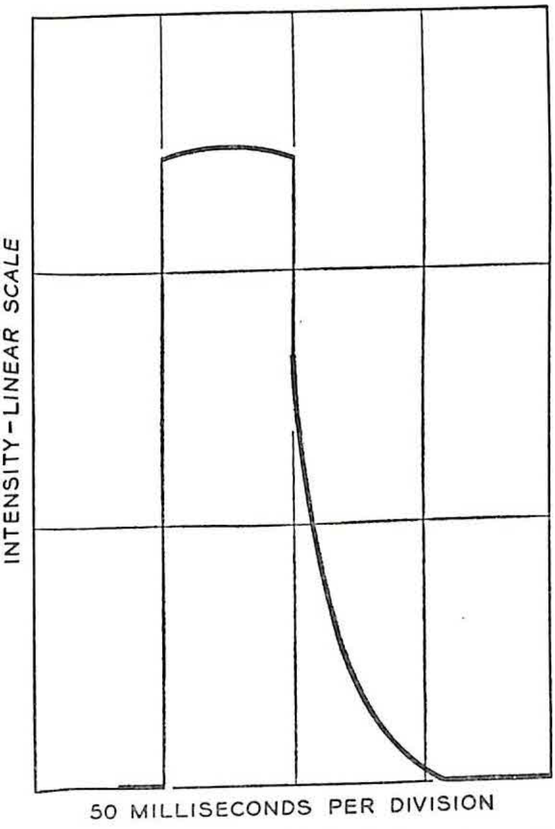


Fig. 1. A 50-millisecond light flash, such as was used in most of the experiments. (Redrawn from a photograph of an oscilloscope trace)

Stimulus materials. The stimuli used in this experiment were lettered 5x8 cards viewed at a distance of 22 inches. The lettering was done with a Leroy No. 5 pen, producing capital letters about 0.45 inch high. Only the 21 consonants were used, to minimize the possibility of *S*s interpreting the arrays as words. In a few sets of cards the letter Y was also omitted. In all, over 500 different stimulus cards were used.

There was very little learning of the stimulus materials either by the *S*s or by the *E*. The only learning that was readily apparent was on several stimuli that had especially striking letter combinations. Except for the stimuli used for training, no *S* ever was required to report the same part of any stimulus more than two or three times, and never in the same session.

Figure 2 illustrates some typical arrays of letters. These arrays may be divided into several categories: (*a*) stimuli with 3, 4, 5, 6, or 7 letters normally spaced on a single line; (*b*) stimuli with six letters closely spaced on a single line (6-massed); (*c*) stimuli having two rows of letters with three letters in each row (3/3), or two rows

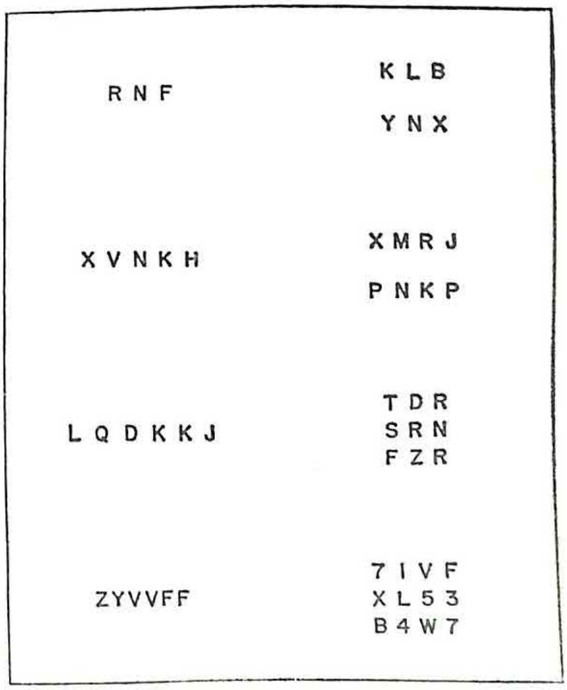


Fig. 2. Typical stimulus materials. Col. 1: 3, 5, 6, 6-massed. Col. 2: 3/3, 4/4, 3/3/3, 4/4/4 L&N.

---

[5] Ralph Gerbrands Company, 96 Ronald Road, Arlington 74, Massachusetts.

of four letters each (4/4); (*d*) stimuli having three rows of letters with three letters in each row (3/3/3). The stimulus information, calculated in bits, for some of the more complex stimuli is 26.4 bits (6-letters, 6-massed, 3/3), 35.1 bits (4/4), and 39.5 bits (3/3/3).

In addition to stimuli that contained only letters, some stimuli that contained both letters and numbers were used. These had eight (4/4 L&N, 35.7 bits) and twelve symbols (4/4/4 L&N, 53.6 bits), respectively, four in each row. Each row had two letters and two numbers—the positions being randomly chosen. The *S* was always given a sample stimulus before L&N stimuli were used and told of the constraint above. He was also told that O when it occurred was the number "zero" and was not considered a letter. Calculated with these constraints, the information in each row of four letters and numbers (17.9 bits) on such a card is nearly equal to the information in a row of four randomly chosen consonants (17.6 bits), even though there are different kinds of alternatives in each case.

Subjects. The nature of the experiments made it more economical to use small numbers of trained *S*s rather than several large groups of untrained *S*s. Four of the five *S*s in the experiment were obtained through the student employment service. The fifth *S* (RNS) was a member of the faculty who was interested in the research. Twelve sessions were regularly scheduled for each *S*, three times weekly.

Instructions and trial procedures. *S* was instructed to look at the fixation cross until it was clearly in focus; then he pressed a button which initiated the presentation after a 0.5-sec. delay. This procedure constituted an approximate behavioral criterion of the degree of dark adaptation prior to the exposure, namely, the ability to focus on the dimly illuminated fixation cross.

Responses were recorded on a specially prepared response grid. A response grid appropriate to each stimulus was supplied. The response grid was placed on the table immediately below the tachistoscope, the room illumination being sufficient to write by. The *S*s were instructed to fill in all the required squares on the response grid and to guess when they were not certain. The *S*s were not permitted to fill in consecutive X's, but were required to guess "different letters." After a response, *S* slid the paper forward under a cover which covered his last response, leaving the next part of the response grid fully in view.

Series of 5 to 20 trials were grouped together without a change in conditions. Whenever conditions or stimulus types were changed, *S* was given two or three sample presentations with the new conditions or stimuli. Within a sequence of trials, *S* set his own rate of responding. The *S*s (except ND) preferred rapid rates. In some conditions, the limiting rate was set by the *E*'s limitations in changing stimuli and instruction tones. This was about three to four stimuli per minute.

Each of the first four and last two sessions began with and/or ended with a simple task: the reporting of all the letters in stimuli of 3, 4, 5, and 6 letters. This procedure was undertaken in addition to the usual runs with these stimuli to determine if there were appreciable learning effects in these tasks during the course of the experiment and if there was an accuracy decrement (fatigue) within individual sessions. Very little improvement was noted after the second session. This observation agrees with previous reports (Whipple, 1914). There was little difference between the beginning and end of sessions.

Scoring and tabulation of results. Every report of all *S*s was scored both for total number of letters in the report which agreed with letters in the stimulus and for the number of letters reported in their correct positions. Since none of the procedures of the experiments had an effect on either of these scores independently of the other, only the second of these, *letters in the correct position*, is tabulated in the results. This score, which takes position into account, is less subject to guessing error,[6] and in some cases it is more readily interpreted than a score which does not take position into account. As the maximum correction for guessing would be about 0.4 letter for the 4/4/4 (12-letter) material—and considerably less for all other materials—no such correction is made in the treatment of the data. In general, data were not tabulated more accurately than 0.1 letter.

Data from the first and second sessions were not used if they fell below an *S*'s average performance on these tasks in subsequent sessions. This occurred for reports of five and of eight (4/4) letters for some *S*s. A similar criterion applied in later sessions for tasks that were initiated later. In this case, the results of the first "training" session(s) are not incorporated in the total tabulation if they lie more than 0.5 letter from *S*'s average in subsequent sessions.

Experiment 1: Immediate-Memory

When an *S* is required to give a complete (whole) report of all the letters on a briefly exposed stimulus, he will generally not report all the letters correctly. The average number of letters which he does report correctly is usually called his *immediate-memory span* or *span of apprehension for that particular stimulus material under the stated observation conditions.* An expression such as immediate-memory span (Miller, 1956a) implies that the number of items reported by *S* remains invariant with changes in stimulating conditions. The present experiment seeks to determine to what extent the span of immediate-memory is independent of the number and spatial arrangement of letters, and of letters and numbers on stimulus cards. If this independence is demonstrated, then the qualification "for that particular stimulus material" may be dropped from the term immediate-memory span when it is used in these experiments.

---

[6] If there are a large number of letters in the stimulus, the probability that these same letters will appear somewhere on the response grid, irrespective of position, becomes very high whether or not *S* has much information about the stimulus. In the limit, the correspondence approaches 100% provided only that the relative frequency of each letter in the response matches its relative frequency of occurrence in the stimulus pack. If the response is scored for both letter *and* position, then the percent guessing correction is independent of changes in stimulus size.

Procedure. *S*s were instructed to write all the letters in the stimulus, guessing when they were not certain. All 12 types of stimulus materials were used. At least 15 trials were conducted with each kind of stimulus with each *S*. Each *S* was given at least 50 trials with the 3/3 (6-letter) stimuli which had yielded the highest memory span in preliminary experiments. The final run made with any kind of stimulus was always a test of immediate-memory. This procedure insured that *S*s were tested for memory when they were maximally experienced with a stimulus.

Results. The lower curves in Fig. 3 represent the average number of letters correctly reported by each *S* for each material.[7] The most striking result is that immediate-memory is constant for each *S*, being nearly independent of the kind of stimulus used. The immediate-memory span for individual *S*s ranges from approximately 3.8 for JC to approximately 5.2 for NJ with an average immediate-memory span for all *S*s of about 4.3 letters. (The upper curves are discussed later.)

The constancy which is characteristic of individual immediate-memory curves of Fig. 3 also appears in the average curve for all *S*s. For example, three kinds of stimuli were used that had six letters each: six letters normally spaced on one line, 6-massed, and 3/3-letters (see Fig. 2). When the data for all *S*s are pooled, the scores for each of these three types of materials are practically the same: the range is 4.1–4.3 letters. The same constancy holds for stimuli containing eight symbols. The average number of letters correctly reported for each of the two different kinds of eight letter stimuli, 4/4, 4/4 L&N, is nearly the same: 4.4, 4.3, respectively.

Most *S*s felt that stimuli containing both letters and numbers were more difficult than those containing letters only. Nevertheless, only NJ showed an objective deficit for the mixed material.

In conclusion, the average number of correct letters contained in an *S*'s whole report of the stimulus is approximately

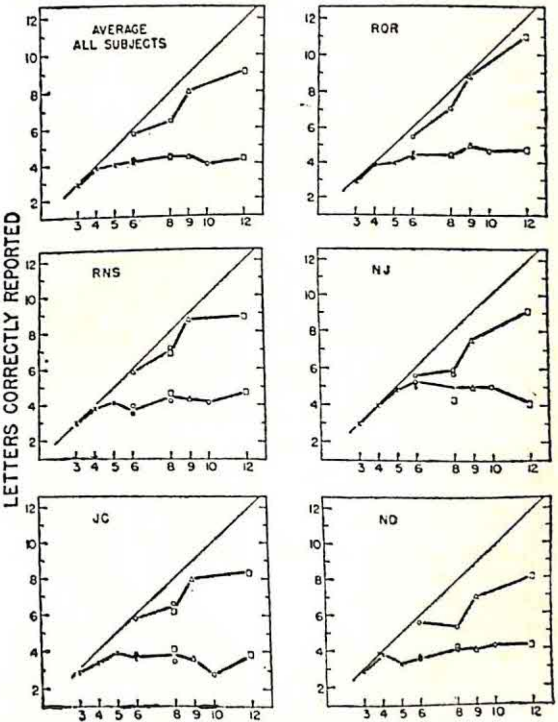


FIG. 3. "Channel capacity curves." Immediate-memory and letters available (output information) as functions of the number of stimulus letters (input information). Lower curves = immediate-memory (Exp. 1); upper curves = letters available immediately after termination of the stimulus; diagonal lines = maximum possible score (i.e., input = output). Code: × = letters on one line; + = 6-massed; o = 3/3, 4/4, 5/5; △ = 3/3/3; □ = 4/4 L&N, 4/4/4 L&N.

---

[7] See Sperling (1959) for tables giving the numerical values of all points appearing in this and in all other figures.

equal to the smaller of *(a)* the number of letters in the stimulus or *(b)* a numerical constant—the span of immediate-memory—which is different for each *S*. The use of the term immediate-memory span is therefore justified within the range of materials studied. This limit on the number of letters that can be correctly reported is an individual characteristic, but it is relatively similar for each of the five *S*s of the study.

Experiment 2: Exposure Duration

The results of Experiment 1 showed that, regardless of material, *S*s could not report more than an average of about 4.5 items per stimulus exposure. In order to determine whether this limitation was a peculiar characteristic of the short exposure duration (0.05 sec.), it was necessary to vary the exposure duration.

Procedure. As in the previous experiment, *S*s were instructed to report all the letters in the stimulus. The stimuli were six letter cards (3/3). NJ, who was able to report more than five correct letters in Experiment 1, was given 4/4 stimuli in order to make a possible increment in his accuracy of responding detectable. The *S*s were given 10 trials in each of the four conditions, .015-, .050-, .150- (.200-), .500-sec. exposure duration, in the order above. In a later session, additional trials were conducted at .015-sec. exposure as a control for Experiment 5. These trials are averaged with the above data.

Results and discussion. Figure 4 illustrates the number of letters correctly reported as a function of the duration of exposure. The main result is that exposure duration, even over a wide range, is not an important parameter in determining the number of letters an *S* can recall correctly. Both individually and as a group, *S*s show no systematic changes in the number of letters correctly reported as the exposure duration was varied from 0.015 to 0.500 sec. The invariance of the number of letters reported as a function of exposure durations up to about 0.25 sec. for the kind of presentation used (dark pre- and post-exposure fields) has long been known (Schumann, 1904).

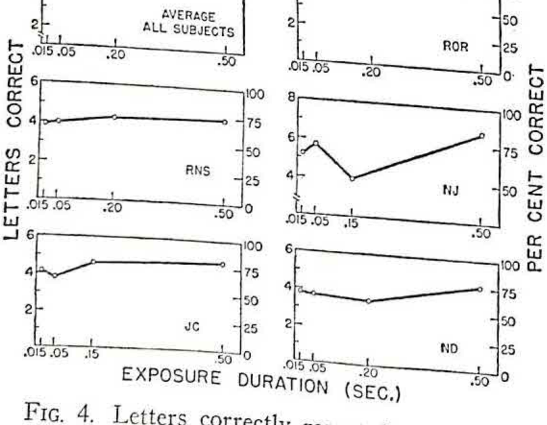


FIG. 4. Letters correctly reported as a function of exposure duration.

Experiment 3: Partial Report

Experiments 1 and 2 have demonstrated the span of immediate-memory as an invariant characteristic of each *S*. In Experiment 3 the principles of testing in a perceptual situation that were advanced in the introduction are applied in order to determine whether *S* has more information available than he can indicate in his limited immediate-memory report.

The *S* is presented with the stimulus as before, but he is required only to make a partial report. The length of this report is four letters or less, so as to lie within *S*'s immediate-memory span. The instruction that indicates which row of the stimulus is to be reported is coded in the form of a tone. The instruction tone is given after the visual presentation. The *S* does not know until he hears the tone which row is called for. This is therefore a procedure which samples the information that *S* has available after the termination of the visual stimulus.

Procedure. Initially, stimulus materials having only two lines were used, that is, 3/3 and 4/4. The *S* was told that a tone would be sounded, that this tone would come approximately simultaneously with the exposure, and that it would be either a high tone (2500 cps) or a low tone (250 cps). If it were a high tone, he was to write only the upper row of the stimulus; if a low tone, only the lower row. He was then shown a sample card of 3/3 letters and given several high and low tones. It was suggested that he keep his eyes fixated on the fixation point and be equally prepared for either tone. It would not be possible to outguess the *E* who would be using a random sequence of tones.

The tone duration was approximately 0.5 sec. The onset of the tone was controlled through the same microswitch that controlled the off-go of the light, with the completion of a connection from an audio-oscillator to the speaker. Intensity of the tone was adjusted so that the high (louder) tone was "loud but not uncomfortable."

In each of the first two sessions, each *S* received 30 training trials with each of the materials 3/3, 4/4. In subsequent sessions *S*s were given series of 10 or more "test" trials. Later, a third, middle (650 cps) tone was introduced to correspond to the middle row of the 3/3/3 and 4/4/4 stimuli. The instructions and procedure were essentially the same as before.

In any given session, each tone might not occur with equal frequency for each type of stimulus. Over several sessions, usually two, this unequal frequency was balanced out so that an *S* had an exactly equal number of high, medium, and low tones for each material. If an *S* "misinterpreted" the tone and wrote the wrong row, he was asked to write what he could remember of the correct row. Only those letters which corresponded to the row indicated by the tone were considered.

Treatment of the Data. In the experiments considered in this section, *S* is never required to report the whole stimulus but only one line of a possible two or three lines. The simplest treatment is to plot the percentage of letters correct. This, in fact, will be done for all later comparisons. The present problem is to find a reasonable measure to enable comparison between the partial report and the immediate-memory data for the same stimuli. The measure, *percent correct,* does not describe the results of the immediate-memory experiments parsimoniously. In Experiment 1 it was shown that *S*s report a constant number of letters, rather than a constant percentage of letters in the stimulus. The measure, *number of letters correct,* is inappropriate to the partial report data because the number of letters which *S* reports is limited by the *E* to at most three or four. The most reasonable procedure is to treat the partial report as a random sample of the letters which the *S* has available. Each partial report represents a typical sample of the number of letters *S* has available for report. For example, if an *S* is correct about 90% of the time when he is reporting three out of nine letters, then he is said to have 90% of the nine letters—about eight letters—available for partial report at the time the instruction tone is given.

In order to calculate the number of available letters, the average number of letters correct in the partial report is multiplied by the number of equiprobable (nonoverlapping), partial reports. If there are two tones and two rows, multiplication is by 2.0; if three, by 3.0. As before, only the number of correct letters in the correct position is considered.

Results. The development of the final, stable form of the behavior is relatively rapid for *S*s giving partial reports. The average for all *S*s after 30 trials (first session) with the 3/3 stimuli was 4.5; on the second day the average of 30 more trials was 5.1. On the third day *S*s averaged 5.6 out of a possible six letters. Most of the improvement was due to just one *S*: ND who improved from 2.9 to 5.8 letters available. In the 3/3/3 stimulus training, all *S*s reached their final value after the initial 40 trials on the first day of training. The considerable experience *S*s had acquired with the partial reporting procedure at this time may account for the quick stabilization. NJ, whose score was 7.7 letters available on the first 20 trials, was given almost 150 additional trials in an unsuccessful attempt to raise this initial score.

In Fig. 3 the number of letters available as a function of the number of letters in the stimulus are graphed as the upper curves. For all stimuli and for all *S*s, the available information calculated from the partial report is greater than that contained in the immediate-memory report. Moreover, from the divergence of the two curves it seems certain that, if still more complex stimuli were available, the amount of available information would continue to increase.

The estimate above is only a lower bound on the number of letters that *S*s have available for report after the termination of the stimulus. An upper bound cannot be obtained from experiments utilizing partial reports, since it may always be argued that, with slightly changed conditions, an improved performance might result. Even the lower-bound measurement of the average available information, however, is twice as great as the immediate-memory span. The immediate-memory span for the 4/4/4 (12-letters and numbers) stimuli ranges from 3.9 to 4.7 symbols for the *S*s, with an average of 4.3. Immediately after an exposure of the 4/4/4 stimulus material, the number of letters available to the *S*s ranged from 8.1 (ND) to 11.0 (ROR), with an average of 9.1 letters available.

This number of letters may be transformed into the bits of information represented by so many letters. For the 4/4/4 (12-letters and numbers) material, the average number of bits available, then, is 40.6, with a range from 36.2 to 49.1 (out of a possible 53.6 bits). These figures are considerably higher than the usual estimates. For example, in a recent review article Quastler (1956) writes: "All experimental studies agree that man can . . . assimilate less than 25 bits per glance" (p. 32). The data obtained in Experiment 3 not only exceed this maximum, but they contain no evidence that the information that became available to the *S*s following the exposure represented a limit of "man" rather than a maximum determined by the limited information contained in the stimuli which were used.

Experiment 4: Decay of Available Information

Part 1: Development of Strategies of Observing

It was established in Experiment 3 that more information is available to the *S*s immediately after termination of the stimulus than they could report. It remains to determine the fate of this surplus information, that is, the "forgetting curve." The partial report technique makes possible the sampling of the available information at the time the instruction signal is given. By delaying the instruction, therefore, decay of the available information as a function of time will be reflected as a corresponding decrease in the accuracy of the report.

Procedure. The principal modification from the preceding experiment is that the signal tone, which indicates to the *S* which row is to be reported, is given at various other times than merely "zero delay" following the stimulus off-go. The following times of indicator tone onset relative to the stimulus were explored: 0.05 sec. before stimulus onset (−0.10 sec.), ±0.0-, +0.15-, +0.30-, +0.50-, +1.0-sec. delays after stimulus off-go. The stimuli used were 3/3, 4/4.

The *S*s were given five or more consecutive trials in each of the above conditions. These trials were always preceded by at least two samples in order to familiarize *S* with the exact time of onset. The particular delay of the instruction tone on any trial was thus fixed rather than chosen randomly. The advantages of this procedure are (*a*) optimal performance is most likely in each delay condition, if *S* is prepared for that precise condition (cf. Klemmer, 1957), (*b*) minimizing delay changes makes possible a higher rate of stimulus presentations. On the other hand, a random sequence of instruction tone delays would make it more likely that *S* was "doing the same thing" in each of the different delay conditions.

The sequence in which the different delay conditions followed each other was chosen either as that given above (ascending series of delay conditions) or in the reverse order (descending series). Within a session, a descending series always followed an ascending series and vice versa, irrespective of the stimulus materials used. At least two ascending and two descending series of delay conditions were run with each *S* and with each material after the initial training (Experiment 3) with that material. This number of trials insures that for each *S* there are at least 20 trials at each delay of the indicator tone.

Results and discussion. The development of the typical behavior is illustrated by the *S*, ROR, in Figs. 5a, b, c. Figure 5a shows ROR's performance in a single session, the first posttraining session. The upper and lower curves represent the ascending and descending series of tone instruction delays, respectively. The arrows indicate the order. Although each point is based upon only five trials, the curves are remarkably similar and regular. Clearly, most of the letters in excess of ROR's memory span are forgotten within about 0.25 sec. The rapid forgetting of these letters justifies calling this a short-term memory and accounts for the fact that it may easily be overlooked under less than optimal conditions. In the following session (Fig. 5b) the descending series

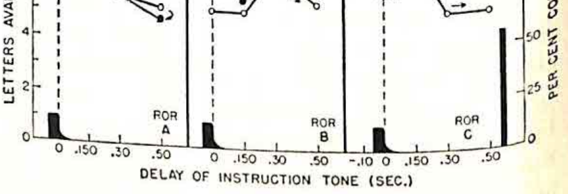


FIG. 5. Partial report of eight (4/4) letters, three consecutive sessions. Arrows indicate the sequence in which conditions followed within a session. The light flash is shown on same time scale at lower left of each figure. Bar at right indicates immediate-memory for this material. One subject (ROR).

was given first. Here orderly behavior disintegrates. In the third session (Fig. 5c) two modifications were introduced: (*a*) the number of trials in each delay condition was increased to eight and (*b*) a new delay condition was given—namely, a signal tone coming 0.05 sec. before the stimulus onset. The curves are again regular, but they are obviously different for the ascending and descending series. For the session indicated in Fig. 5c, an analysis of the errors by position shows that in the ascending series the errors are evenly split between the top and bottom rows of the stimulus; in the descending series, the top row is favored 3:1.

ROR's performance is analyzable in terms of two kinds of observing behavior (strategies) which the situation suggests. He may follow the instruction, given by *E* prior to training, that he pay equal attention to each row. In this case, errors are evenly distributed between rows. Or, he may try to anticipate the signal by guessing which instruction tone will be presented. In this case, *S* is differentially prepared to report one row. If the signal and *S*'s guess coincide, *S* reports accurately; if not, poorly. Such a guessing procedure would lead to the variability observed in Fig. 5b. On the other hand, *S* may prefer always to anticipate the same row—in the case of ROR (Fig. 5c, descending series), the top row. This would again allow reliable scores, provided only that there are an equal number of instruction signals calling for the top and bottom rows. Concomitantly, a differential accuracy of report for the two rows is observed. (ROR's preference for the top row is again prominent in Experiment 7, Figs. 10, 11.)

Equal attention responding is initially reinstated on the third day. The obvious change in procedure which is responsible is the introduction of a tone 0.05 sec. before the stimulus onset (−0.10 sec. "after" its termination). This signal is sufficiently in advance of the stimulus so that perfect responding is possible by looking at only the row indicated by the signal tone. ROR scores 100%, both in this condition and in the succeeding zero delay. The whole (ascending series) decay curve of Fig. 5c is highly similar to that of Fig. 5a. A run with 3/3 stimuli was interposed between the ascending and descending series shown in Fig. 5c. Since the guessing procedures were easily sufficient for a nearly perfect score with 3/3 materials, when the descending series of delay conditions was run, ROR continued guessing. While guessing was advantageous at the long delays, at the short delays it was a decided disadvantage.

Figure 6 illustrates the performances of RNS, a sophisticated observer. RNS described the two strategies (equal attention, guessing) to *E*. In accordance with the instructions to do as well as he could, RNS said that he switched from the first to the second strategy at delays longer than 0.15 sec. Thus in the three short delay conditions, RNS divided his errors almost evenly (19:21) between the favored (top) and the unfavored rows; in the two longer delays, errors were split 4:26. The dip in the curve indicates that RNS did not switch strategies quite as soon as he could have, for optimal performance. Such a dip is

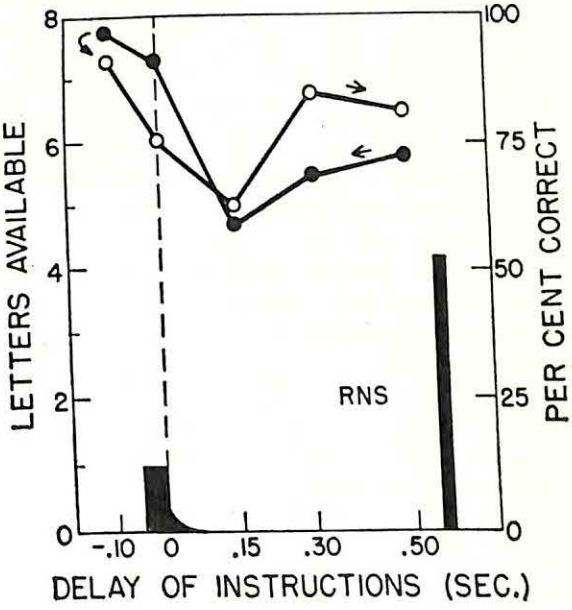


FIG. 6. Partial report of eight (4/4) letters, last of three sessions. Arrows indicate the sequence in which conditions followed within session. The light flash is shown on same time scale at lower left. Bar at right indicates immediate-memory for this material. (RNS)

characteristically seen in experiments of this kind.

The other *S*s exhibit similar curves.[8] These are not presented, as the main features have already been demonstrated by ROR and RNS. In summary, the method of delaying the instruction tone is a feasible one for determining the decay of the short-term memory contents, but experience with the difficult, long delays causes a considerable increase in the variability of *S*s' performance which is carried over even to the short delay conditions. This has been attributed to *S*s' change from an equal attention to a guessing strategy in observing the stimulus.[9]

Part 2: Final Level of Performance

The analysis of the preceding experiment has indicated that two distinct kinds of observing behavior develop when the instruction to report is delayed. The accuracy of report resulting from the first of these behaviors (equal attention) is correlated with the delay of the instruction tone; it is associated with the *S*s initially giving equal attention to all parts of the stimulus. The accuracy of the other kind of report (guessing) is uncorrelated with the delay of the instruction; it is characterized by *S*s' differential preparedness for some part of the stimulus (guessing). Equal attention observing is selected for further study here. The preceding experiment suggests three modifications that would tend to make equal attention observing more likely to occur, with a corresponding exclusion of guessing.

1. The use of stimuli with a larger number of letters, that is, 3/3/3 and 4/4/4. Differential attention to a constant small part of the stimulus is less likely to be reinforced, the larger the stimulus. The use of three tones instead of two diminishes the probability of guessing the correct tone.

2. Training with instruction tones that begin slightly before the onset of the stimulus. It is not necessary for *S* to guess in this situation since he can succeed by depending upon the instruction tone alone. This situation not only makes equal attention likely to occur, but differentially reinforces it when it does occur.

When delays of longer length are tested, priority should be given to an ascending series of delays so that *S* will, at the beginning of a particular delay sequence, have a high probability of entering with the desired observing behavior. This probability might be nearly 1.0 by interposing a series of trials on which the instruction is given in advance of, or immediately upon, termination of the stimulus and requiring that *S* perform perfectly on this task before he can continue to the particular delay being tested.[10] This tedious procedure was tried; but, as it did not have an appreciable effect upon the results, it was discontinued. The problem is that *S*s learn to switch between the two modes of behavior in a small number of trials.

3. The *E* may be able to gain *verbal* control over *Ss'* modes of responding. Initially, however, even *S* cannot control his own behavior exactly. This suggests a limit to what *E* can do. For example, frequently *S*s reported that, although they had tried to be equally prepared for each row, after some tones they realized that they had been selectively prepared for a particular row. This comment was made both when the tone and the row coincided and (more frequently) when they differed.

---

[8] See Sperling (1959) for tables containing individual averages of all trials for each *S* at each delay of instruction tone: 3/3, 4/4.

[9] Increase in variability (with consequent decrement in accuracy and/or speed) is not unusual after difficult conditions. For example, Cohen and Thomas (1957) in a clinically oriented study have reported an exactly analogous "hysteresis" phenomenon in a study of discriminative reaction time. Hysteresis refers to the fact that, when the difficulty of an experimental task is changed, the corresponding change in accuracy of response lags behind the change in task.

[10] It takes, on the average, a very large number of trials for *S* to get 10 consecutive perfect trials even if he has 6 or 7 of 9 letters available or knows with 2/3 probability what the tone will be. Success in this task within a reasonable time limit demands a level of excellence reached only with "equal attention" observing, as judged by the other criteria.

Some verbal control is, of course, possible. An instruction that was well understood was:

> You will see letters illuminated by a flash that quickly fades out. This is a visual test of your ability to read letters under these conditions, not a test of your memory. You will hear a tone during the flash or while it is fading which will indicate which letters you are to attempt to read. Do not read the card until you hear the tone, [etc.].

The instruction was changed at the midway point in the experiment. The *S* was no longer to do as well as he could by any means, but was limited to the procedure described above. Part 2 of this experiment, utilizing 9- and 12-letter stimuli, was carried out with the three modifications suggested above.

Results. The results for 3/3/3 and 4/4/4 letters and numbers are shown for each individual *S* in Figs. 7 and 8. The two ordinates are linearly related by the equation:

$$\frac{\text{percent correct}}{100} \times \text{no. letters in stimulus} = \text{letters available}$$

Each point is based on all the test trials in the delay condition. The points at zero delay of instructions for NJ and JC also include the training trials, as these *S*s showed no subsequent improvement.

The data indicate that, for all *S*s, the period of about one sec. is a critical one for the presentation of the instruction to report. If *S*s receive the instruction 0.05 sec. before the exposure, then they give accurate reports: 91% and 82% of the letters given in the report are correct for the 9- and 12-letter materials, respectively. These partial reports may be interpreted to indicate that the *S*s have, on the average, 8.2 of 9 and 9.8 of 12 letters available. However, if the instruction is delayed until one sec. after the exposure, then the accuracy of the report drops 32% (to 69%) for the 9-letter stimuli, and 44% (to 38%) for the 12-letter stimuli. This substantial decline in accuracy brings the number of

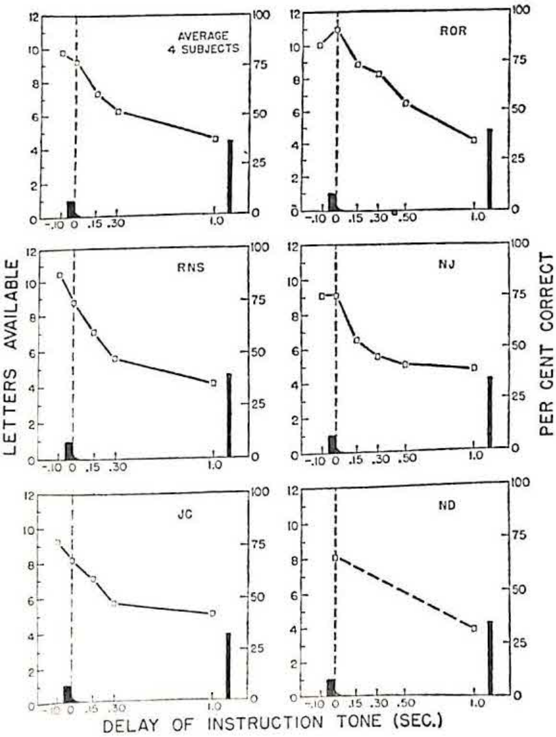


FIG. 7. Decay of available information: nine (3/3/3) letters. Light flash is shown on same time scale at lower left. Bar at right indicates immediate-memory for this material.

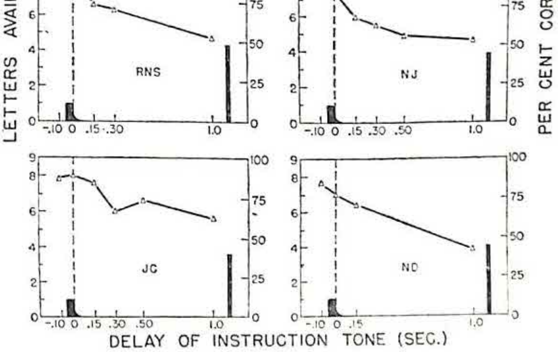


FIG. 8. Decay of available information: twelve (4/4/4) letters and numbers. Light flash is shown on same time scale at lower left. Bar at right indicates immediate-memory for this material.

letters available very near to the number of letters that *S*s give in immediate-memory (whole) reports.

The decay curves are similar and regular for each *S* and for the average of all *S*s. Although individual differences are readily apparent, they are small relative to the effects of the delay of the instruction. For example, when an instruction was given with zero delay after the termination of the stimulus, the *least* accurate reports by any *S*s are given by ND, who has 8.1 letters available immediately after the termination of the stimulus. With a one-sec. delay of instructions, the *most* accurate reports were given by JC, who has only 5.1 letters available at this time.

In Fig. 3, in which whole reports and partial reports were compared, only that particular partial report was considered in which the instruction tone followed the stimulus with zero delay. It is evident from this experiment that the zero delay instruction is unique only in that it is the earliest possible "after" instruction, but not because of any functional difference.

In Fig. 9, therefore, the 0.15-, 0.50-, and 1.00-sec. instruction delays are also plotted for one *S*, ROR. Data for the six- (3/3) and eight- (4/4) letter stimuli are taken from the two ascending delay series with each material that yielded monotonic results. Figure 9 clearly highlights the significance of a precisely controlled coded instruction, given within a second of the stimulus off-go, for the comparison of partial and immediate-memory reports. One second after termination of the stimulus, the accuracy of ROR's partial reports is no longer very different from the accuracy of his whole reports.

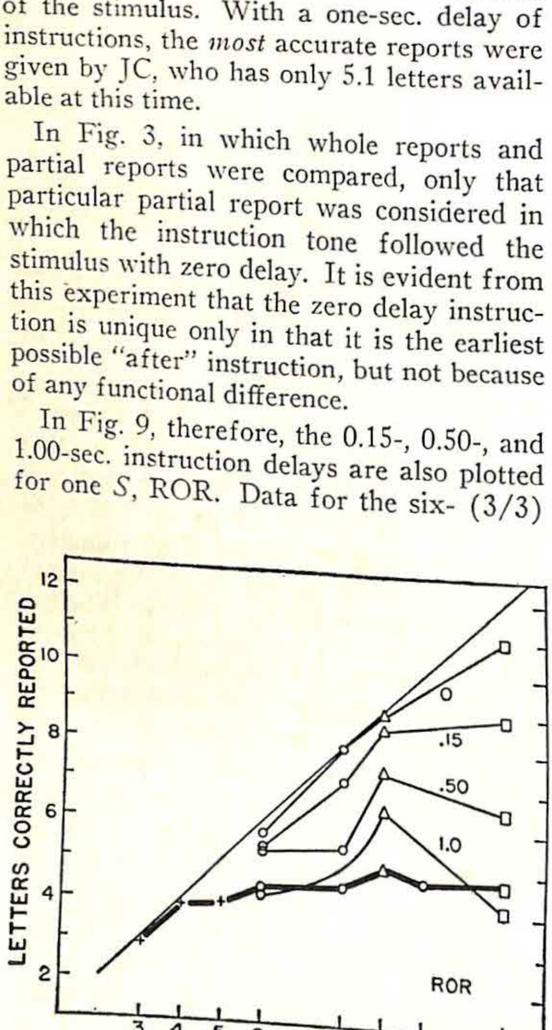


FIG. 9. Immediate-memory and available information. The parameter is the time at which available information is sampled (delay of instruction). Heavy line indicates immediate-memory for the same materials. One subject (ROR).

Experiment 5: Some Exposure Parameters

In Experiment 3 it was shown that the number of letters reported correctly is almost independent of the exposure duration over a range from 15 to 500 msec. It is well known, however, that the relation between the accuracy of report and the exposure duration depends upon the pre- and post-exposure fields (Wundt, 1899).

In a technique developed in Helmholtz's laboratory (Baxt, 1871) the informational (stimulus) field is followed, after a variable delay, by a noninformational, homogeneous, bright post-exposure field. Using this method, Baxt showed that the number of reportable letters was a nearly linear function of the delay of the bright post-exposure field.[11]

Other combinations of pre- and post-exposure fields have also been tried (Dodge, 1907a). The usual tachistoscopic presentation utilizes gray pre- and post-exposure fields (Woodworth, 1938). Baxt's procedure, however, is the most disadvantageous for the observer. A similar procedure was therefore selected, in order to study whole and partial reports in a vastly different visual presentation from that of the previous experiments.

---

[11] This important method was described by Ladd (1899) and James (1890) in their textbooks, but it is no longer well known. Consequently it has been "rediscovered," most recently, by Lindsley and Emmons (1958). Baxt (1871) intended that the bright second field would interfere with the lingering image of the first (informational) field. Unfortunately, the effect depends in a complex way upon the intensity of the two fields. Derived time values must be used with caution. In some cases the Baxt technique may actually result in no loss of legibility, the second field producing a negative "afterimage" instead of merely interfering with the positive image (cf. Footnote 15).

Procedure. 1. *S*s were instructed to write all the letters of the stimulus; 3/3 stimuli were used. After several sample presentations of a stimulus card followed by a light post-exposure field, *S*s were given a random sequence of normal (pre- and post-exposure fields dark) and Baxt (pre-exposure dark, post-exposure field light) trials. The Baxt trials do not correspond exactly to the presentation that Baxt used. In this experiment, the post-exposure field is the same intensity as the stimulus (informational) field, whereas Baxt usually used more intense post-exposure fields; also, the stimulus field always remains *on* until the onset of the post-exposure field, whereas Baxt used a fixed five-msec. duration for the stimulus field. The post-exposure field itself remains on for about one sec. The pre-exposure field is always dark, as in all the previous experiments. Two exposure durations, 0.015 and 0.050 sec., were tested.

2. Three *S*s were tested with the Baxt presentation of a 3/3/3 stimulus at an exposure duration of 0.015 sec. The partial report procedure was used to determine the effects of the post-exposure field on the number of letters available.

3. The same three *S*s were run as their own controls. The procedure was exactly the same as in Paragraph 2 above except that the post-exposure field was normal (dark).

Results. The complete results are given in Tables 1 and 2. In all tests, the Baxt procedure reduces the response accuracy of all *S*s to about one-half of their normal score. This finding confirms the earlier studies. However, a linear relation between exposure duration and the number of letters reported was not observed. The failure to find a linear relation may be due to the previously mentioned differences between the presentations.

For RNS and ND, the number of letters available is nearly the same (about two) in the partial report of 9-letter stimuli as the whole report of 6-letter stimuli. The fact that in both procedures the number of letters given by *S*s is the same suggests that a Baxt presentation reduces the number, or the length of time that letters are available, and that it does not directly affect the immediate-memory span.

ROR's partial reports of Baxt presentations are considerably more accurate than those of the other *S*s, although they are not as accurate as his reports in control presentations. ROR seemed to show improvement on successive Baxt trials. JC, another *S* who seemed to show improvement, was given additional Baxt trials on which he continued to improve slowly. Unfortunately, it was unfeasible to determine the asymptotic performance of these two *S*s. Whether the difference in performance between ROR, and RNS and ND is attributable to some overt response, such as squinting or blinking, was not determined.

TABLE 1

A Comparison of Response Accuracy with Two Different Post-Exposure Fields

Subject	Exposure (sec.)	Normal	Baxt
RNS	(0.015)	3.9	2.5
	(0.050)	4.0	2.2
ROR	(0.015)	4.8	3.5
	(0.050)	4.4	2.8
ND	(0.015)	3.8	1.9
	(0.050)	3.7	2.3
NJ	(0.015)	5.1	2.4
	(0.050)	5.4	3.4
JC	(0.015)	4.1	2.7
	(0.050)	3.8	3.4
Mean	(0.015)	4.3	2.6
	(0.050)	4.3	2.8

Note.—Number of letters correctly reported. Whole report of six (3/3) letter stimuli. Normal = pre- and post-exposure fields dark; Baxt = pre-exposure field dark, post-exposure field white.

Table 2 also enables the comparison of 0.015- and 0.050-sec. exposures of 3/3/3 stimuli. A decrease in exposure duration has only a slight effect on the number of letters available. This suggests, as in the immediate-memory experiments, that the duration of a tachistoscopic exposure is not as important a determinant of the number of letters available as the fields which follow the exposure.

Experiment 6: Letters and Numbers

In Experiment 3 partial reports were found to be uniformly more accurate than

TABLE 2

A Comparison of Response Accuracy with Two Different Post-Exposure Fields

Delay of Instruction (sec.):	0.0		0.15	
Exposure Duration (sec.): Subject	0.015	0.05	0.015	0.05
RNS (N)	8.0	8.7	5.4	6.6
(B)	2.0		2.2	
ROR (N)	8.6	8.9	8.3	8.5
(B)	6.3		5.4	
ND (N)	7.3	7.0	5.8	6.4
(B)	2.2		1.7	
Mean (N)	8.0	8.2	6.5	7.2
(B)	3.5		3.1	

Note.—Number of letters available (fraction of letters correct in partial report × number of letters in stimulus). Stimuli: nine (3/3/3) letters. (N) = normal (pre- and post-exposure fields dark). (B) = Baxt (pre-exposure field dark, post-exposure field white).

whole reports. In one case, stimuli of eight letters were used and only one row of four letters was reported. Designating the letters to be reported by their location is only one of a number of possible ways. In the following experiment, a quite similar set of stimuli is used; each stimulus has two letters and two numbers in each of the two rows. The partial report again consists of only four symbols, but these are designated either as letters or as numbers rather than by row. In addition, a number of controls which are also relevant to Experiment 3 are conducted.

Procedure. I. Training: The *S*s were given practice trials with the instruction: "Write down only the numbers if you hear a short pip (tone 0.05-sec. duration) and only the letters if you hear the long tone (0.50-sec. duration)." The tones were then given with zero delay following the stimulus off-go. The stimuli were 4/4 L&N.

II. In the following session, tests were conducted with five different instructions:

1. Letters only—Instructions given well in advance of stimulus to write only the letters in the following card(s). (8 trials)

2. Numbers only—Write only the numbers in the following card(s). (8 trials)

3. Top only—Write only the top row in the following card(s). (4 trials)

4. Bottom only—Write only the bottom row in the following card(s). (4 trials)

5. Instruction tone—Write either letters or numbers as indicated by tone. Tone onset 0.05 sec. before stimulus onset. (16 trials) ROR was also given additional trials at longer delay times.

Results. The results are illustrated in Table 3. For purposes of comparison, the number of correct letters is multiplied by two when an instruction was used which required *S* to report only four of the eight symbols of the stimulus. This includes instructions given well in advance of the stimulus. All measures, then, have 8.0 as the top score and are thus equivalent within a scale factor to percent correct measures. The range is 0-8 instead of 0-100. Scores which are based on partial reports are therefore directly comparable to the partial report scores (letters available) obtained in Experiments 3, 4, and 5.

When stimuli consist of letters and numbers, but *S*s report only the letters or only the numbers, then the *S*s' partial reports are only negligibly more accurate than their whole reports of the same stimuli. The average number of letters available (calculated from the partial report) is just 0.2 letter above the immediate-memory span

TABLE 3

Comparison of Five Procedures

Subject	Letters only	Numbers only	Average L&N	Instr. tone −0.10	Immediate-memory	One row only
RNS	5.0	4.5	4.8	4.3	4.6	7.3
ROR	6.5	6.5	6.5	6.3	4.5	7.3
ND	3.5	3.8	3.6	4.1	4.1	7.5
NJ	4.0	5.0	4.5	4.6	4.3	—
JC	3.3	4.0	3.6	3.4	4.1	8.0
Mean	4.5	4.8	4.6	4.5	4.3	—

Note.—Average letters and/or numbers available (fraction of letters—numbers—correct in partial report × number of symbols in stimulus). Stimuli: eight (4/4) letters and numbers.

for the same material. For practical purposes, the partial report score is the same score that *S*s would obtain if they wrote all the letters and numbers they could (that is, gave a whole report) but were scored only for letters or only for numbers, independently by the experimenter. The partial report of letters only (or of numbers only) does not improve even when the instruction is given long in advance verbally instead of immediately before the exposure by a coded signal tone.

The estimate of the number of available letters and numbers which is obtained from the partial report of letters (or numbers) only is also the same as the estimate that would be obtained if, on each trial, *S*s wrote only one row—either the top or the bottom—according to their whim. Reporting only one row of four letters and numbers is a task at which the *S*s succeed with over 90% accuracy. Even if they are scored for the whole stimulus, by arbitrarily reporting only one row they would still achieve a score of almost 50% correct or almost four letters available. This is why no delay series were conducted. If *S*s had ignored the instruction to write only the letters (or numbers) and had written only a single row on each trial, they would have shown less than a 0.5 letter decrement, no matter what the delay of the instruction.

Only ROR showed a substantial improvement when reporting only the numbers (or letters). He was the only *S* with whom it made sense to conduct a systematic delay series, although checks with other *S*s confirmed this conclusion. Table 4 indicates that two extra symbols are available to ROR for report only when the tone is given before the stimulus, but not if it is given immediately after. It should be noted that the information in the instruction tone comes only after it has been on for 0.05 sec. At this time it either continues or is terminated. The actual "instruction" is thus given 0.05 sec. after the tone onset. ROR therefore requires that the *instruction* be given within 0.05 sec. of the stimulus termination if any benefit of the partial report procedure is to be retained.

Whether the *S*s would have shown improvement with a large amount of additional training in the partial report of letters or numbers cannot be stated. Table 3 shows that, when *S*s are required in advance to report only one row, this task is trivial. The substantial advantage of partial reports of rows (report by position) over partial re-

TABLE 4

Partial Reports of Letters or Numbers

Subj.	Prior Verbal Instr.	Delay of Instr.		Tone (sec.)	Immediate Memory
		−0.10	0.0	+0.25	
ROR	6.5	6.3	4.7	4.4	4.5

Note.—Average of symbols available (fraction of letters—numbers—correct in partial report × number of symbols in stimulus). Stimuli: eight (4/4) letters and numbers.

ports of numbers or letters (report by category) when the instruction is given verbally long in advance of the exposure is retained even when the instruction is coded and given shortly after the exposure.

The failure in Experiment 6 to detect a substantial difference in accuracy between partial reports of only letters (or only numbers) and whole reports clearly illustrates that partial reports by position are more effective for studying the capacity of short-term information storage than partial reports by category.

Experiment 7: Order of Report

Interpretations of the effects of instructions upon the report following a single brief visual exposure have often been concerned with either the perceptual sensitizing effects of an instruction given before the exposure or with the importance of forgetting between the exposure and a post-exposure instruction to report. The decay curves of Experiment 4 include both of these effects. Previous studies, however, have usually assumed the order in which the various parts (aspects, dimensions, etc.) of the stimulus are reported to be the significant correlate of post-exposure forgetting. The possibility that information might be well retained even though not immediately reported has been mentioned (Broadbent, 1958), but experimental investigations of such an effect by an independent variation of the order of report by Wilcocks (1925), Lawrence and Laberge (1956), and Broadbent (1957a) have apparently shown otherwise. Broadbent (1957a) has also shown a case in which independent variation of the order of report did not reduce overall response accuracy.

In the present experiment, order of report is introduced as a purely "nuisance" variable for the S. The S is instructed to get as many letters correct as possible, but the E randomly manipulates the order in which they are to be reported. The experiment is a survey of how Ss adapt to this kind of interference with the normal order of their report.

Procedure. The Ss were instructed to write the row indicated by the tone (high, low) first, then to write the other row. They were to try to get as many *total* letters correct as possible, it being of no importance in which particular row the correct letters might be. The instruction tone was given with 0.0-, .30- (or .50-), and 1.0-sec. delay after the termination of the stimulus.

Controls. In addition to the trials with a high or low tone, two sets of 8 (or 10) trials were given with a neutral, middle tone. The instruction was: "Write all the letters in any order you wish, but do not begin writing until you hear the tone." The tone was sounded with 0.0-sec. delay following termination of the stimulus and also with 1.0-sec. delay. It bears repeating here that Ss were not permitted just to mark X's but were required to guess various letters.

Results. Controls: The instruction which required Ss to wait for 1.0 sec. before beginning to write their answer was ignored by the Ss, since it was almost physically impossible to begin writing sooner. Consequently the two different controls—trials on which S was required to "wait" for 1.0 sec. and trials on which S could begin his report immediately—are grouped together. These data, which are almost exactly the same as the memory span data (Experiment 1), are presented on the far right in Fig. 10.

The Ss' responses on the control trials are analyzed in terms of the correlation between the location of letters on the stimulus and the accuracy of the report of these letters in the response. The symbols T and B above I-M in Fig. 10 represent the percentage of the letters of the top and of the bottom rows that Ss report correctly. The middle point is the average percentage of the letters of the top and bottom rows that were correctly reported by Ss. The middle point is therefore also the average percentage correct of all the letters that were reported. Figure 10 shows that all Ss report the top row of the stimulus more accurately than the bottom row, if they are not instructed with regard to the order in which they must report the rows.

The average accuracy with which Ss report the top and the bottom rows, when instructions to report one or the other of these rows are given with various delays after the exposure, is also illustrated in

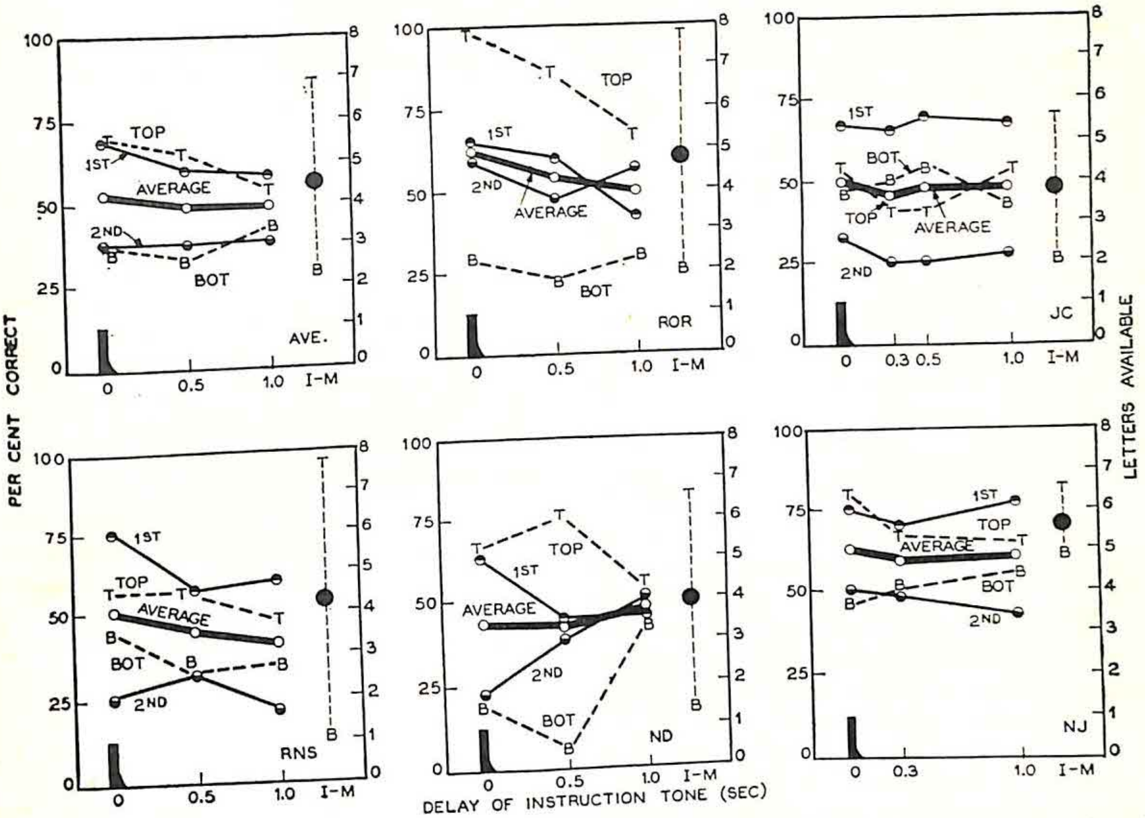


Fig. 10. Accuracy of the first (second) row reported and of the top (bottom) row as a function of the delay of the instruction to report one row first. Light flash is indicated at lower left; immediate-memory (I-M) at right.

Fig. 10. The accuracy of reports of the top row decreases slightly as the delay of the instructions increases. In other respects, however, the data show no systematic changes in accuracy with changes in the delay of instructions. The data clearly indicate that the top row is generally reported more accurately than the bottom row although the instruction to report each row is given with equal frequency.

The same data may also be analyzed with regard to the accuracy of the row that must be reported first and the row that is reported last. All *S*s except ROR are more accurate when they report the first row (the row called for by the instruction tone) than the second row. For most *S*s, therefore, the order of report is correlated with the accuracy of report.

There is a slight tendency for the accuracy of report of the row which is reported first to decrease as the delay of the instructions increases. On the whole, however, the overall accuracy of report decreases slightly with the delay of the instruction to report one row first. The experimental interference with the normal order of report does not change the overall number of letters reported correctly by any *S* by more than about 0.5 letter.

In this task, unlike the preceding ones, individual differences are more striking than the similarities. The pooled data are highly untypical of three of the five *S*s. Figure 11 was devised as a two-dimensional, graphical analysis of variance to compress the details of Fig. 10 into one figure.[12] Each coordinate represents the accuracy of

---

[12] Figure 11 is based upon a suggestion by E. B. Newman. A statistical analysis of variance was not attempted since it would have had to be carried out separately on each *S*. There was not enough data to make this worthwhile, and Fig. 11 serves the same purpose.

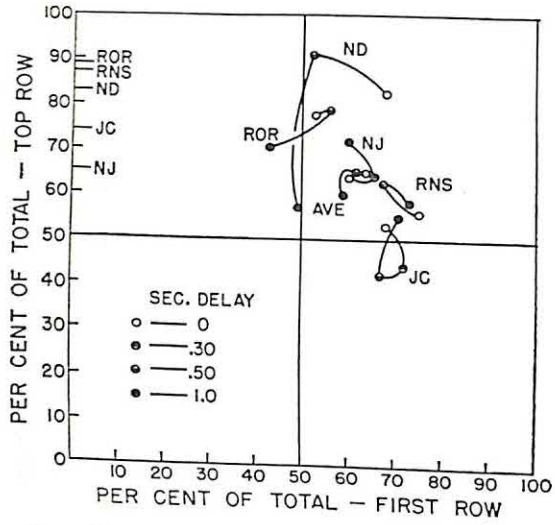


FIG. 11. Graphical analysis of position on stimulus vs. order of report as contributors to response accuracy. Each point represents the average of all trials of an *S* at a particular delay of instruction. The order in which the points are connected corresponds to the magnitude of the delays. Upper left: position preference in control (immediate-memory) report.

one row of the report relative to the whole report. Thus the ordinate represents the number of letters that an individual *S* reports correctly in the top row of the stimulus (independently of order) divided by the total number of letters (both rows) that he reports correctly. Similarly, the abscissa represents the percentage of the total correct letters reported by *S* that are contained in his report of the first row. Since each coordinate is relative to *S*'s own accuracy, no point of the graph is inaccessible to *S* provided that, if necessary, he is willing to sacrifice some accuracy. Since the interference with *S*'s order of report in this experiment had only slight effects on the overall accuracy of *S*'s report, this method of presenting the data is justifiable.

From Fig. 11 it is immediately evident that, for example, 50% of the correct letters that JC reports are from the top row and, by implication, 50% are from the bottom row of the stimulus. More than 70% of the correct letters that JC reports are in the first row reported by him. ROR represents the converse, preferring to report the top row accurately, remaining indifferent to whether it is called for first or last. Other *S*s lie between these extremes, each *S* maintaining approximately the same relative accuracy for the top and the first rows throughout the various delay conditions. Each *S*, therefore, operates within a characteristic, limited area of the graph. ND is an exception. At zero delay of the instruction tone, both position and order account heavily for the correct letters reported by ND. At 0.5-sec. delay, ND ignored the order (preferring to concentrate on the top row), and at 1.0-sec. delay she lost her position (top row) preference as well. In these three conditions, the total number of letters correctly reported by ND remained approximately the same, within 0.5 letter of the control condition. At 1.0-sec. delay *neither position nor the order contribute to ND's accuracy of report.*

All *S*s operate in the upper right quadrant of Fig. 11. This illustrates the finding that no *S* consistently reported the bottom row more accurately than the top, nor the last row better than the row first called for by the instruction tone. It does not, of course, indicate that *S*s could not report the bottom row or the last row more accurately under other conditions. While *S*s normally behave quite consistently, the data of ND show that they may try a number of different procedures. The instructions given the *S*s prior to the experiment were not restrictive. No specific procedure for making a report was suggested to the *S*s, because the purpose of the experiment was to find out how *S*s respond when they are not given detailed instructions. With suitable instructions, training, and reinforcements, *S*s could probably be induced to make most of the possible kinds of reports that can be diagramed by Fig. 11. This remains an empirical problem.

The results obtained in this experiment support the conclusions that both a position preference and the order of report ordinarily correlate with the accuracy of response, but that probably *neither are necessary* conditions for response accuracy. Some *S*s can relinquish the position preference and a favorable order of report with no appreciable decrement in accuracy. This finding is in opposition to Lawrence and Laberge's (1956) contention that accuracy is accounted for by the order of report. Accuracy and order are often correlated, but if a favorable order of report is not *necessary* for accuracy, then it cannot be the cause of accuracy.

When *S* is given a signal indicating which row is to be reported first (Experiment 7), the accuracy of report of the row indicated by the signal (the first row reported) may be compared to the accuracy of the partial report (Experiment 3). The overt procedure on each trial is quite similar in Experiments 3 and 7. The only difference is that in the order of report experiment, after the *S*s have finished writing the row indicated by the signal, they must also write down the other row. In the partial report procedure they do not have to write the second row. The partial report and the order of report experiments also share a common dependent variable: the accuracy of report of the row indicated by the instruction signal.

In view of the similarity in procedure, it is surprising that the accuracy of this common datum should be so different in the two experiments. For example, when *S*s give only the partial reports (the instruction signal being given immediately after termination of the stimulus), then they report 90% of the letters correctly in one row of 4/4 stimuli. When they are required to write the other row also, then they report only 69% of the first four letters correctly. Every *S*, individually, gives a more accurate partial report (Experiment 3) than a report of the first row—of two rows to be reported (Experiment 7). The consistent superiority of the partial report over the first half of a whole report prevails even when the instruction to report is delayed for 0.5 (or 0.3) sec. In all cases where data are available, each *S* reports a row of four letters more accurately when he does not have to write another row of four letters afterwards. That what *S*s must write later should affect the accuracy of what they write first must be explained—if we disregard teleological explanation—by the effect of prior instructions on the accuracy of the report. In other words, if order of report is effective in determining the accuracy of report, then this effect must be a function of instructions given prior to any report at all. For some *S*s, no effect of order of report upon response accuracy was observed.

The two findings, that partial reports are uniformly more accurate than whole reports and that order of report may be uncorrelated with accuracy, contradict Lawrence and Laberge's conclusion that "partial" reports are essentially similar to "first" reports. In fact, the second finding (that in some circumstances order of report and accuracy of report are not correlated) provides a direct counterexample to their conclusion. Their different results may be in part due to the vastly different stimuli which they used. Lawrence and Laberge's entire stimuli each contained less information than two randomly chosen letters.

DISCUSSION

In all seven experiments, *S*s were required to report the letters of briefly exposed lettered stimuli. Two kinds of reports were explored: *partial reports,* which required the *S*s to report only a specified part of the stimulus, and *whole reports,* which required the *S*s to report all the letters of the stimulus. Experiment 3 demonstrated that the accuracy of partial reports was consistently greater than the accuracy of whole reports. Another important difference between partial and whole reports is the correlation of accuracy with the delay of the instruction to report. This was shown in Experiment 4 in which the time delay of the instruction signal, which indicated the row of the stimulus to be reported, was varied. The accuracy of the partial report was found to be a sharply decreasing function of the time at which the instruction was given. If the instruction signal was delayed for one sec. after the exposure, the accuracy of the partial report was no longer very different from that of the whole report. In Experiment 7 it was shown that the accuracy of the whole report does not change as the time of the signal to report is varied—over the same range of time—up to one sec. after the exposure.

The two kinds of report can also be considered in terms of the information (in these experiments, letters) which they indicate the *S* has available for report. In the whole report, the *S* reports all the information that he can. When he gives a partial report, the *S* may have additional available information that is not required for the report. A calculation of the information available to the *S*s for their partial reports indicates that between two and three times more information is available for partial reports than for the whole reports. This discrepancy between the two kinds of report is short-lived. Information in excess of that indicated by the whole report was available to the *S*s for only a fraction of a second following the exposure. At the end of this time, the accuracy of partial reports is no longer very different from that of whole reports.

The whole report has already been extensively studied by psychologists. The maximum number of items an individual can give in such a report is called his *span of immediate-memory;* whole reports are usually called *immediate-memory reports.* Experiments 1 and 2 extend the well-known conclusions that the span of immediate-memory is an individual characteristic and that it is constant over a wide range of stimuli and exposure conditions. Although, in immediate-memory experiments, items are conventionally presented sequentially, Experiments 1 and 2 illustrate that this is not necessary—that a simultaneous presentation may also give results characteristic of immediate-memory experiments.

The main problems to be considered here concern the partial–whole report discrepancy: (*a*) Why is the partial report more accurate than the whole report? (*b*) Why does the partial report retain this added accuracy only for a fraction of a second after the exposure?

The answers proposed are a systematic elaboration of an observation that is readily made by most viewers of the actual tachistoscopic presentation. They report that the stimulus field appears to be still readable at the time a tone is heard which follows the termination of the stimulus by 150 msec. In other words, the subjective image or sensation induced by the light flash outlasts the physical stimulus at least until the tone is heard. The stimulus information is thus "stored" for a fraction of a second as a persisting image of the objective stimulus. As the visual image fades, its legibility (information content) decreases, and consequently the accuracy of reports based upon it decreases.

There is other evidence, besides such phenomenological accounts, that suggests that information is available in the form of an image for a short time after extinction of the physical stimulus. In the first place, it is inconceivable that the observers should stop seeing the stimulus at exactly the moment the light is turned off. The rise and fall of sensation may be rapid, but they are not instantaneous. The question is not *whether* the observer continues to see the stimulus after the illumination is turned off, but for *how long* he continues to see the stimulus. A number of different kinds of psychophysical measurements of the rise and fall of sensation have been attempted. These estimates of the persistence of the visual sensation vary from a minimum of 0.05 sec. (Wundt, 1899) to almost one sec. (McDougall, 1904). The most representative estimates are in the neighborhood of 1/6 sec. (cf. Piéron, 1934), a figure that is in good agreement with the results.[13]

---

[13] Measurements of the persistence of sensation have almost invariably used techniques which have at most questionable validity. Wundt's method depends upon masking, the effect of the persisting stimulus upon another stimulus. The masking power of a stimulus may be quite different from its visibility. McDougall's measurements, as well as those cited by Piéron, depend upon motion of a stimulus across the retina. Such measurements are undoubtedly influenced by the strong temporal and spatial interactions of the eye (Alpern, 1953). Schumann's ingenious application of the method of Baxt to the determination of persistence is probably the only experiment that utilizes pattern stimulation. The other methods have not been tried with pattern stimuli although there is, a priori, no good reason why they have not been. The possibility that the persistence of pattern information is quite different from persistence of "brightness" has not been investigated.

In Experiment 5 it was shown that the post-exposure field strongly influences the accuracy of both the partial and the whole report. This experiment indicates that the available information is sensitive to interference by *noninformational visual stimuli* which follow the exposure. The dependence of available information upon noninformational visual stimulation is just the dependence that would be expected of a visual image.

Finally, there are subtle aspects of the sequence of letters reported by an *S* which characterize the information that is available to him. In sequentially spoken letters, for example, there is a limit—two—on the number of letters that can be adjacent to any given letter. Different limits apply to a two-dimensional visual display. If information is stored in a form topologically similar to the stimulus, this may be detected by noting the sequential dependencies that limit successive responses to the stimulus.

Probably the kinds of sequential responding that would most clearly distinguish visual from auditory information storage would be (*a*) the ability of the *S* to read the rows of the visual stimulus backwards as well as forwards, or to report the columns or the diagonals, and (*b*) his inability to do an equivalent task when presented with the information sequentially. (All these procedures merely require the report of adjacent letters if the stimulus is two dimensional.) Unfortunately, these particular experiments were not conducted.

The foregoing experiments offer some relevant evidence. In contrast to spoken letters and numbers, which are most accurately recalled if they occur at the beginning or end of a sequence (Pollack, 1952),[14] no obvious gradients of accuracy were found in the foregoing experiments. The middle row actually tended to be slightly better reported than the other rows. Therefore, it is unlikely that the *entire* visual stimulus (12-letters and numbers) was transformed into an auditory (sequential) representation for storage. Such an entire transformation is also unlikely, though not impossible, because of the relatively small time between the stimulus exposure and the report.

An analysis of errors reveals numerous cases of errors that may be classified as "misreading" (for example, confusions between E and F, B and R) and as "mishearing" (for example, confusions between B and D, D and T—Miller & Nicely, 1955). Still other confusions (for example, C and G) are ambiguous. All of these types of errors occurred whenever errors occurred at all. The ubiquity of misreading and mishearing errors, taken at face value, suggests that both visual and auditory storage of information are always involved in both whole and partial reports. A nonquantitative error analysis is therefore not likely to shed much light on the question of visual imagery. The frequent mishearing errors suggest that the storage of letters, just prior to a written report, may share some of the characteristics of audition. Like the preceding analysis of the constraints upon successive responses, error analysis requires considerable research before it can be quantitatively applied to problems of imagery.

This then is the evidence—phenomenological reports, the effects of the post-exposure fields, the known facts of the persistence of sensation, and the detailed characteristics of the responses—that is consistent with the hypothesis that information is initially stored as a visual image and that the *S*s can effectively utilize this information in their partial reports. In the present context, the term, visual image, is taken to mean that (*a*) the observer behaves as though the physical stimulus were still present when it is not (that is, after it has been removed) and that (*b*) his behavior in the absence of the stimulus remains a function of the same variables of visual stimulation as it is in its presence. The units of a visual image so defined are always those of an equivalent "objective image," the physical stimulus. It is as logical or illogical to compute the information contained in a visual image (as was done in Experiments 3 and 4) as it is to compute the information in a visual stimulus.

---

[14] Summarized in Luce (1956).

"Visual image" and "persistence of sensation" are terms suggested by the asynchrony between the time during which a stimulus is present and the time during which the observer behaves as though it were present. Although asynchrony is inevitable for short exposure durations, there is, of course, no need to use the term "visual image" in a description of this situation. One might, for example, refer simply to an "information storage" with the characteristics that were experimentally observed. This form of psychological isolationism does injustice to the vast amount of relevant researches.

Imagery that reputedly occurs long after the original stimulation (memory images, eidetic images, etc.) is of interest as well as imagery that occurs for only a few tenths of a second following stimulation. Whether the term "imagery," as it has been used here to describe the immediate effects of brief stimulation, is an appropriate term for the description of the lasting effects of stimulation is an empirical problem. It is hoped, however, that the principles and methods developed here will not be without relevance to these traditional problems.

Persistence of Vision and Afterimages

Between the short persistence of vision and the remembrance of a long-passed event, there is an intermediate situation, the afterimage, which requires consideration. In discussing afterimages, it will be useful to distinguish some phases of vision that normally follow an intense or prolonged stimulus. First, there is the "initial" (or primary, or original) "image" (or sensation, or impression, or perception, or response). Any combination of a term from the first and from the second of these groups may be used. The initial image is followed by a latent period during which nothing is seen and which may in turn be followed by a complex sequence of afterimages. Afterimages may be either positive or negative; almost any sequence is possible, but the initial image is almost always positive.[15] Some authors distinguish the initial image from a positive afterimage (for example, McDougall, 1904); others do not (for example, von Helmholtz, 1924-25). It is often implicit in such distinctions that the persistence of the initial image is due to a continued excitatory process, whereas afterimages arise from receptor fatigue. If there is no repeated waxing and waning of sensation, but merely a single rise and fall, one cannot distinguish two phases in the primary image, one corresponding to the "initial image" and the other to an identical "positive afterimage" of it.

Although it is difficult to prove that visual information is stored in the initial image, there can be no gainsaying that an afterimage may be a rich store of information. Positive or negative afterimages may carry many fine details, including details that were not visible at the time of stimulation (von Helmholtz, 1924-25). Afterimages generally last for at least several seconds, and following high energy stimulation they normally last for several minutes (Berry & Imus, 1935). The clarity of the details, of course, deteriorates with the passage of time. Since afterimages appear to move when the eye is moved, they usually have been considered retinal phenomena. Taken together, these facts imply that *there is a considerable capacity for visual information storage in the retina.* If the illumination of the stimulus cards used in the foregoing experiments had been sufficiently intense to blaze the letters upon the retina and thereby take maximum advantage of its information storage capacity, there would have been little doubt afterwards as to the nature or location of most of the available information. The stimulus presentations actually used, however, rarely elicited reports of afterimages; *S*s usually reported seeing simply a single brief flash. The problem is therefore to determine the persistence of the image of the brief flash, or equivalently, the duration of seeing (the stimulus) or the persistence of vision (of the stimulus), rather than the duration of an afterimage. These terms are used to suggest that the *S* feels he is responding directly to the stimulus rather than to aftereffects of stimulation.

---

[15] Bidwell (1897) and Sperling (1960) describe conditions for seeing a negative "after-image without prior positive image." The method involves a presentation quite similar to that of Baxt (1871).

Psychologists have often carelessly assumed that the absence of discernible afterimages following a visual presentation was sufficient to insure that the duration of sensation will correspond to the duration of the physical stimulus, that is, that there is no persistence of vision at all. Wundt (1899) was one of the first to take vigorous exception to this naive view. Wundt's most compelling example was drawn from Weyer (1899). Weyer had found that two 40-microsecond light flashes had to be separated by 40 to 50 msec. in order for them to be seen as two distinct flashes; at smaller separations they were seen as a single flash. In the dark adapted eye, the minimum separable interval that consistently yielded reports of "two flashes" was 80 to 100 msec.

Wundt argued that the two flashes could not be seen as distinct until the sensation occasioned by the first flash had ceased. Thus, under optimum conditions, the minimum duration of the sensation of a short flash was at least 40 msec. which, in this case, was 1,000 times the stimulus duration. Wundt thought that, in order to determine the duration of a longer flash, one must merely add the 40 msec. of fade-out time to the actual physical light duration. While these details of Wundt's reasoning may be questioned, his main point, based on the example of the short flash, is indisputable: one does not directly control the time for which information is visually available simply by manipulating exposure duration. The experiments reported here provide a direct proof of this assertion.

An Application of the Results to "Before and After" Experiments

The previous experiments showed that more information is available to *S*s for a few tenths of a second after the exposure than they can give in a complete report of what they have seen. It was suggested that the limit on the number of items in the memory report is a very general one, the span of immediate-memory, which is relatively independent of the nature of the stimulus. Evidence was offered that information in excess of the immediate-memory span is available to the *S* as a rapidly fading visual image of the stimulus. If more information is available to him than he can remember, the *S* must "choose" a part of it to remember. In doing so, he has chosen the part to forget. In Experiments 3 and 4, *S*s exercised only locational choices, that is, portions of the stimulus were remembered only on the basis of their location. Locational choices are probably not the only effective choices that the *S* can make. During the short time that information is available to him, the *S* may process it in any way in which he normally handles information. Usually, what he does, or attempts to do, is determined by the instructions. The *S*'s (unobservable) response to the stimulus is probably the same whether the instruction to make this response is given before the stimulus presentation or after it; the difference between the two cases lies in the information that the response can draw upon. If the stimulus contains more information than the *S*'s immediate-memory span, and if the post-exposure instruction is delayed until the *S* has little of this extra information available, then a difference in the accuracy of the responses with prior- and post-exposure instructions will be observed. If the stimulus does not contain more information than can be coded for immediate-memory, or if the post-exposure instruction is given soon enough so that the *S* can utilize the still available information effectively, then only minor differences in the accuracy of responses with prior- and post-exposure instructions will be observed. If the stimulus is destitute of information (for example, a single, mutilated, dimly illuminated letter of the alphabet) then a host of other factors which are normally insignificant may become crucial. In this case, the "stimulus" itself may well be irrelevant (Goldiamond, 1957), and the effects of instructions given before or after the exposure must be predicted on some other basis.

There are some simple experiments in which it is known a priori that the effects of instructions given either before or after the exposure will be exactly the same. This degenerate situation can be illustrated by a stimulus which is exposed for one microsecond and with sufficient energy to be clearly visible. By suitable coding, the pre-exposure and post-exposure instructions can be separated by only two microseconds. The example serves to emphasize that what is implicitly referred to by "before and after" is not the exposure but something else: traditionally, the sensitization and/or forgetting that presumably occur in conjunction with the exposure. Thus, the theory that has been presented here merely gives an explicit statement of assumptions that have long been implicit.

Unobservable Responses and the Order of Report

The subjective response to the high signal tone is "looking up." Since eye movements cannot occur in time to change the retinal image with any of the presentations used (Diefendorf & Dodge, 1908) a successful looking-up must be described in terms of a shift in "attention." Nonetheless, such a shift in attention can be quantitatively studied by means of a stabilized retinal image (Pritchard, 1958) although Wundt (1912), who did not use this modern, technically difficult technique, was able to give many essential details. The reaction time for the attentional response, like the reaction time for more observable responses, is greater than zero. Therefore, if the *S* is given an instruction before the presentation, he can prepare for, or sensitize himself to, the correct row of the stimulus even though there is not time enough for a useful eye movement. The response to an instruction which is given 0.05 sec. before the stimulus is probably the same as the response to a similar instruction that is given 0.1 sec. later, immediately after the exposure. The short time difference, 0.1 sec., accounts for the similar accuracy of responding in these two conditions.

Once his attention is directed to the appropriate row, the *S* still has to read the letters. This, too, takes time. Baxt's (1871) data indicate that the time required to read a letter is about 10 msec. Baxt's experiment, with some modifications, was repeated by the author, and similar results were obtained.[16]

How is all this relevant to the order of report? The order in which the letters are finally reported can be an important variable because of (*a*) purely temporal factors (letters that are reported first will be more accurately reported only because they are reported sooner after the exposure, the actual order of reporting the letters per se being relatively unimportant) or (*b*) interaction effects (the report of some letters is detrimental to the report of the remaining letters, that is, letters reported later suffer from proactive interference by the letters reported earlier).

That purely temporal factors cannot be very important can be seen from the slope of the curves describing available information as a function of time. In the foregoing experiments, the amount of available information approached the immediate-memory span at about 0.5 to 1.0 sec. after the exposure; further decrements in available information as a function of time are slight. The report of the letters usually does not begin until 1.0–1.5 sec. after the exposure. The passage of time during the actual time that letters are being reported, therefore, cannot account for appreciable accuracy changes as a function of the order of report.

The second possible effect of order of report—the interfering effect of the letters reported first upon unreported letters—cannot be so readily discounted. Proactive interference would imply that partial reports are more accurate than whole reports by an amount dependent upon their relative lengths. The results of Experiment 4 tend to support this view. At delays of the instruction signal greater than 0.3 sec., partial reports of three letters (from stimuli of nine letters) indicate more available letters than do partial reports of four letters (from stimuli of twelve letters). On the other hand, in Experiment 7 one *S*, ROR, does not show decreased accuracy as a function of the length of his report. ROR is able to report eight letters as accurately when he begins his report with three or four incorrect letters as when he ends his report with three or four incorrect letters. Other *S*s did not systematically attempt to report incorrect letters first. Had they been required to report incorrect letters first, they might well have been able to do so.

---

[16] Sperling, G. Unpublished experiments conducted at the Bell Telephone Laboratories, 1958.

The choice of what part of the stimulus to attend to or of which letters to read is the choice of what fraction of the stimulus information to utilize. This choice can be made successfully only while the information is still available. The *S*s prefer to report what they remember first, but this does not imply that they remember it because they report it first. It is difficult to disentangle the many factors that determine precisely what stimulus letters will appear in the response, but important choices of what information is to be recalled must occur while there is still something to choose from. Since the actual report begins only when there are available but a few letters in excess of the immediate-memory span, the order of report can at most play only the minor role of determining which of the few "excess" letters will be forgotten.[17]

The Questions of Generality and Repeatability

To what extent are the results obtained limited to the particular conditions of the foregoing experiments? The possibility that the actual physical fading of the light source is important to the availability of information can be rejected not only on prior grounds (see *Apparatus*) but also by the empirical findings. For example, in Fig. 9, the curve representing the number of letters available 0.15 sec. after exposure is quite similar to the 0.0-sec. delay curve. There is no visible energy emitted by the light source 0.15 sec. after its termination.

In the present case, the answer to the problem of repeatability of the results is made less speculative by three separate investigations that have since been conducted with similar techniques to those reported here.

The experiments have been repeated by the author[18] with a different tachistoscope, timer, and a light source that has only a negligible afterglow. All the main findings were reproducible.

Klemmer and Loftus (1958) confirmed the existence of a short-term, high information storage. They used a display consisting of four discrete line patterns, the *S* being required to report only one of these. The instruction was coded either as a signal light or verbally. Decay curves obtained when the instruction is delayed are similar to those reported above. A similar experiment has also been conducted by Averbach,[19] who used a television tachistoscope to present stimuli containing up to 16 letters. A pointer appeared above the letter to be reported. Initially, *S*s had about twelve letters available for report, but the number decreased rapidly when the visual instruction was delayed.

It is usually technically more difficult to code instructions visually than acoustically. Although the principle of sampling in order to determine available information is common to both kinds of instructions, visually coded instructions differ in some interesting ways from acoustically coded instructions. For example, the time taken to "interpret" —or even to find—a visual instruction may well depend on its location relative to the fixation point. Moreover, there may be spatial interactions between the visual "instruction" and the symbols to be reported. On the other hand, prior to training, the task of interpreting a visual marker is easier for *S*s than the equivalent task with an acoustically coded instruction. Ultimately, such differences are probably only of secondary importance since the two kinds of experiments agree quite well.

---

[17] There are many ways in which proactive interference might occur. For example, if letters are stored "sequentially" prior to report (cf. Broadbent, 1957b), then the importance of order of report may lie in the agreement of the two sequences: storage and report.

[18] Sperling, G. Unpublished experiments conducted at the Bell Telephone Laboratories, 1958.

[19] Averbach, E. Unpublished experiments conducted at the Bell Telephone Laboratories, 1959.

Three main findings emerge from the experiments reported here: a large amount of information becomes available to observers of a brief visual presentation, this information decays rapidly, the final level is approximately equal to the span of immediate-memory. Although the exact, quantitative aspects of information that becomes available following a brief exposure unquestionably depend upon the precise conditions of presentation, it seems fair to conclude that the main results can be duplicated even under vastly different circumstances in different laboratories.

SUMMARY AND CONCLUSIONS

When stimuli consisting of a number of items are shown briefly to an observer, only a limited number of the items can be correctly reported. This number defines the so-called "span of immediate-memory." The fact that observers commonly assert that they can *see* more than they can *report* suggests that memory sets a limit on a process that is otherwise rich in available information. In the present studies, a sampling procedure (partial report) was used to circumvent the limitation imposed by immediate-memory and thereby to show that at the time of exposure, and for a few tenths of a second thereafter, observers have two or three times as much information available as they can later report. The availability of this information declines rapidly, and within one second after the exposure the available information no longer exceeds the memory span.

Short-term information storage has been tentatively identified with the persistence of sensation that generally follows any brief, intense stimulation. In this case, the persistence is that of a rapidly fading, visual image of the stimulus. Evidence in support of this hypothesis of visual information storage was found in introspective accounts, in the type of dependence of the accuracy of partial reports upon the visual stimulation, and in an analysis of certain response characteristics. These and related problems were explored in a series of seven experiments.

An attempt was first made to show that the span of immediate-memory remains relatively invariant under a wide range of conditions. Five practiced observers were shown stimuli consisting of arrays of symbols that varied in number, arrangement, and composition (letters alone, or letters and numbers together). It was found (Experiments 1 and 2) that each observer was able to report only a limited number of symbols (for example, letters) correctly. For exposure durations from 15 to 500 msec., the average was slightly over four letters; stimuli having four or fewer letters were reported correctly nearly 100% of the time.

In order to circumvent the immediate-memory limit on the (whole) report of what has been seen, observers were required to report only a part—designated by location—of stimuli exposed for 50 msec. (partial report). The part to be reported, usually one out of three rows of letters, was small enough (three to four letters) to lie within the memory span. A tonal signal (high, middle, or low frequency) was used to indicate which of the rows was to be reported. The *S* did not know which signal to expect, and the indicator signal was not given until after the visual stimulus had been turned off. In this manner, the information available to the *S* was sampled immediately after termination of the stimulus.

Each observer, for each material tested (6, 8, 9, 12 symbols), gave partial reports that were more accurate than whole reports for the same material. For example, following the exposure of stimuli consisting of 12 symbols, 76% of the letters called for in the partial report were given correctly by the observers. This accuracy indicates that the total information available from which an observer can draw his partial report is about 9.1 letters (76% of 12 letters). This number of randomly chosen letters is equivalent to 40.6 bits of information, which is considerably more information than previous experimental estimates have suggested can become available in a brief exposure. Furthermore, it seems probable that the 40-bit information capacity observed in these experiments was limited by the small amount of information in the stimuli rather than by a capacity of the observers.

In order to determine how the available information decreases with time, the instruction signal, which indicated the row of the stimulus to be reported, was delayed by various amounts, up to 1.0 sec. (Experiment 4). The accuracy of the partial report was shown to be a sharply decreasing function of the delay in the instruction signal. Since, at a delay of 1.0 sec., the accuracy of the partial reports approached that of the whole reports, it follows that the information in excess of the immediate-memory span is available for less than a second. In contrast to the partial report, the accuracy of the whole report is not a function of the time at which the signal to report is given (Experiment 7).

The large amount of information in excess of the immediate-memory span, and the short time during which this information is available, suggests that it may be stored as a persistence of the sensation resulting from the visual stimulus. In order to explore further this possibility of visual information storage, some parameters of visual stimulation were studied. A decrease of the exposure duration from 50 to 15 msec. did not substantially affect the accuracy of partial reports (Experiment 5). On the other hand, the substitution of a white post-exposure field for the dark field ordinarily used greatly reduced the accuracy of both partial and whole reports. The ability of a homogeneous visual stimulus to affect the available information is evidence that the process depends on a persisting visual image of the stimulus.

Whether other kinds of partial reports give similar estimates of the amount of available information was examined by asking observers to report by category rather than by location. The observer reported numbers only (or the letters only) from stimuli consisting of both letters and numbers (Experiment 6). These partial reports were no more accurate than (whole) reports of all the letters and numbers. The ability of observers to give highly accurate partial reports of letters designated by location (Experiment 3), and their inability to give partial reports of comparable accuracy when the symbols to be reported are designated as either letters or numbers, clearly indicates that all kinds of partial reports are not equally suitable for demonstrating the ability of observers to retain large amounts of information for short time periods.

In the final study (Experiment 7), the order of report was systematically varied. Observers were instructed to get as many letters correct as possible, but the order in which they were to report the letters was not indicated until after the exposure. An instruction tone, following the exposure, indicated which of the two rows of letters on the stimulus was to be reported first. This interference with the normal order of report reduced only slightly the total number of letters that were reported correctly. As might be expected, the first row—the row indicated by the instruction tone—was reported more accurately than the second row (order effect). There was, however, a strong tendency for the top row to be reported more accurately than the bottom row (position effect). Although, as a group, the observers showed both effects, some failed to show either the order or the position effect, or both. The fact that, for some observers, order and position are not correlated with response accuracy suggests that order of report, and position, are not the major *causes* of, nor the necessary conditions for, response accuracy. The high accuracy of partial report observed in the experiments does not depend on the order of report or on the position of letters on the stimulus, but rather it is shown to depend on the ability of the observer to read a visual image that persists for a fraction of a second after the stimulus has been turned off.

REFERENCES

Adams, J. S. The relative effectiveness of pre- and post-stimulus setting as a function of stimulus uncertainty. Unpublished master's dissertation, Department of Psychology, University of North Carolina, 1955.

Alpern, M. Metacontrast. *J. Opt. Soc. Amer.*, 1953, **43**, 648–657.

Baxt, N. Über die Zeit welche nötig ist damit ein Gesichtseindruck zum Bewusstsein kommt und über die Grösse (Extension) der bewussten Wahrnehmung bei einem Gesichtseindrucke von gegebener Dauer. *Pflüger's Arch. ges. Physiol.*, 1871, **4**, 325–336.

Berry, W., & Imus, H. Quantitative aspects of the flight of colors. *Amer. J. Psychol.*, 1935, **47**, 449–457.

Bidwell, S. On the negative after-images following brief retinal excitation. *Proc. Roy. Soc. Lond.*, 1897, **61**, 268–271.

Bridgin, R. L. A tachistoscopic study of the differentiation of perception. *Psychol. Monogr.*, 1933, **44**(1, Whole No. 197), 153–166.

Broadbent, D. E. Immediate memory and simultaneous stimuli. *Quart. J. exp. Psychol.*, 1957, **9**, 1–11. (a)

Broadbent, D. E. A mechanical model for human attention and memory. *Psychol. Rev.*, 1957, **64**, 205–215. (b)

Broadbent, D. E. *Perception and communication.* New York: Pergamon, 1958.

Cattell, J. McK. Über die Trägheit der Netzhaut und des Sehcentrums. *Phil. Stud.*, 1883, **3**, 94–127.

Chapman, D. W. The comparative effects of determinate and indeterminate aufgaben. Unpublished doctor's dissertation, Harvard Uniiversity, 1930.

Chapman, D. W. Relative effects of determinate and indeterminate "Aufgaben." *Amer. J. Psychol.*, 1932, **44**, 163–174.

Cohen, L. D., & Thomas, D. R. Decision and motor components of reaction time as a function of anxiety level and task complexity. *Amer. Psychologist*, 1957, **12**, 420. (Abstract).

Dallenbach, K. M. Attributive vs. cognitive clearness. *J. exp. Psychol.*, 1920, **3**, 183–230.

Diefendorf, A. R., & Dodge, R. An experimental study of the ocular reactions of the insane from photographic records. *Brain*, 1908, **31**, 451–489.

Dodge, R. An experimental study of visual fixation. *Psychol. Monogr.*, 1907, **8**(4, Whole No. 35). (a)

Dodge, R. An improved exposure apparatus. *Psychol. Bull.*, 1907, **4**, 10–13. (b)

Erdmann, B., & Dodge, R. *Psychologische Untersuchungen über das Lesen auf experimenteller Grundlage.* Halle: Niemeyer, 1898.

Glanville, A. D., & Dallenbach, K. M. The range of attention. *Amer. J. Psychol.*, 1929, **41**, 207–236.

Goldiamond, I. Operant analysis of perceptual behavior. Paper read at Symposium on Experimental Analysis of Behavior, APA Annual Convention, 1957.

James, W. *The principles of psychology.* New York: Holt, 1890.

Klemmer, E. T. Simple reaction time as a function of time uncertainty. *J. exp. Psychol.*, 1957, **54**, 195–200.

Klemmer, E. T., & Loftus, J. P. *Numerals, nonsense forms, and information.* USAF Cambridge Research Center, Operational Applications Laboratory, Bolling Air Force Base, 1958. (Astia Doc. No. AD110063)

Külpe, O. Versuche über Abstraktion. In, *Bericht über den erste Kongress für experimentelle Psychologie.* Leipzig: Barth, 1904. Pp. 56–68.

Ladd, G. T. *Elements of physiological psychology: A treatise of the activities and nature of the mind.* New York: Scribner, 1889.

Lawrence, D. H., & Coles, G. R. Accuracy of recognition with alternatives before and after the stimulus. *J. exp. Psychol.*, 1954, **47**, 208–214.

Lawrence, D. H., & LaBerge, D. L. Relationship between accuracy and order of reporting stimulus dimensions. *J. exp. Psychol.*, 1956, **51**, 12–18.

Lindsley, D. B., & Emmons, W. H. Perception time and evoked potentials. *Science*, 1958, **127**, 1061.

Long, E. R., Henneman, R. H., & Reid, L. S. Theoretical considerations and exploratory investigation of "set" as response restriction: The first of a series of reports on "set" as a determiner of perceptual responses. *USAF WADC tech. Rep.*, 1953, No. 53–311.

Long, E. R., & Lee, W. A. The influence of specific stimulus cuing on location responses: The third of a series of reports on "set" as a determiner of perceptual responses. *USAF WADC tech. Rep.*, 1953, No. 53–314. (a)

Long, E. R., & Lee, W. A. The role of spatial cuing as a response-limiter for location responses: The second of a series of reports on "set" as a determiner of perceptual responses. *USAF WADC tech. Rep.*, 1953, No. 53–312. (b)

Long, E. R., Reid, L. D., & Garvey, W. D. The role of stimulus ambiguity and degree of response restriction in the recognition of distorted letter patterns: The fourth of a series of reports on "set" as a determiner of perceptual responses. *USAF WADC tech. Rep.*, 1954, No. 54-147.

Luce, D. R. *A survey of the theory of selective information and some of its behavioral applications.* New York: Bureau of Applied Social Research, 1956.

McDougall, W. The sensations excited by a single momentary stimulation of the eye. *Brit. J. Psychol.*, 1904, **1,** 78–113.

Miller, G. A. Human memory and the storage of information. *IRE Trans. Information Theory*, 1956, **IT–2,** No. 3, 129–137. (a)

Miller, G. A. The magic number seven, plus or minus two: Some limits on our capacity for processing information. *Psychol. Rev.*, 1956, **63,** 81–97. (b)

Miller, G. A., & Nicely, P. E. An analysis of perceptual confusions among some English consonants. *J. Acoust. Soc. Amer.*, 1955, 27, 338–352.

Piéron, H. L'évanouissement de la sensation lumineuse: Persistance indifferenciable et persistance totale. *Ann. psychol.*, 1934, **35,** 1–49.

Pollack, I. The assimilation of sequentially-encoded information. *Hum. Resources Res. Lab. memo Rep.*, 1952, No. 25.

Pritchard, R. M. Visual illusions viewed as stabilized retinal images. *Quart. J. exp. Psychol.*, 1958, **10,** 77–81.

Quastler, H. Studies of human channel capacity. In H. Quastler, *Three survey papers.* Urbana, Ill.: Control Systems Laboratory, Univer. Illinois, 1956. Pp. 13–33.

Schumann, F. Die Erkennung von Buchstaben und Worten bei momentaner Beleuchtigung. In, *Bericht über den erste Kongress für experimentelle Psychologie.* Leipzig: Barth, 1904. Pp. 34–40.

Schumann, F. The Erkennungsurteil. *Z. Psychol.*, 1922, **88,** 205–224.

Sperling, G. Information available in a brief visual presentation. Unpublished doctor's dissertation, Department of Psychology, Harvard University, 1959.

Sperling, G. Afterimage without prior image. *Science*, 1960, **131,** 1613–1614.

von Helmholtz, H. *Treatise on physiological optics.* Vol. II. *The sensations of vision.* (Transl. from 3rd German ed.) Rochester, New York: Optical Society of America, 1924–25.

Wagner, J. Experimentelle Beitrage zur Psychologie des Lesens. *Z. Psychol.*, 1918, **80,** 1–75.

Weyer, E. M. The Zeitschwellen gleichartiger und disparater Sinneseindrucke. *Phil. Stud.*, 1899, **15,** 68–138.

Whipple, G. M. *Manual of physical and mental tests.* Vol. I. *Simpler processes.* Baltimore: Warwick & York, 1914.

Wilcocks, R. W. An examination of Külpe's experiments on abstraction. *Amer. J. Psychol.*, 1925, **36,** 324–340.

Woodworth, R. S. *Experimental psychology.* New York: Holt, 1938.

Wundt, W. Zur Kritik tachistosckopischer Versuche. *Phil. Stud.* 1899, **15,** 287–317.

Wundt, W. *An introduction to psychology.* London: Allen & Unwin, 1912.

(Early publication received April 28, 1960.)

Vol. 74, No. 12 Whole No. 499, 1960

Psychological Monographs: General and Applied

THE USE OF THEMATIC APPERCEPTION TO ASSESS MOTIVATION IN A NATIONWIDE INTERVIEW STUDY[1]

JOSEPH VEROFF, JOHN W. ATKINSON, SHEILA C. FELD, AND GERALD GURIN
Survey Research Center, University of Michigan

IN recent years, a number of important explorations of the use of projective measures in the survey interview setting (Campbell, 1950; Douvan & Walker, 1956; Lansing & Heyns, 1959; Maccoby & Maccoby, 1954; Sanford, 1950; Sanford & Rosenstock, 1952; Scott, 1956; Tompkins & Miner, 1958) have foreshadowed an inevitable direction of future research in social psychology: the common use in survey studies of assessment techniques developed in experimental and clinical studies of personality dynamics. This trend can be expected to promote a more meaningful integration than now exists of the societal data continually amassed in surveys of broad populations and the research findings accumulated in the clinics and laboratories of personality research. How quickly this integration will come about will depend in large measure upon the conscientiousness with which certain serious methodological problems are faced by the investigator who leaps from the laboratory or clinic to the doorstep of a startled housewife, interviewed because she happened to be selected by the magic of probability sampling. Conceptual integration of facts from field study, clinic, and experiment will be hastened if the personality assessment devices selected from the laboratory have already been investigated with some thoroughness.

The present investigation is an attempt to foster such an integration. We present here an examination of the methodological problems faced in the first attempt to introduce thematic apperceptive measures of three motives (n Achievement, n Affiliation, and n Power) into a nationwide sample survey. Considerable research has been conducted to develop these measures in the experimental setting (Atkinson, 1958; McClelland, Atkinson, Clark, & Lowell, 1953). Consequently, in adapting these measures to the survey setting, we are in a position to be guided by a fund of research findings which define their validity and highlight the methodological issues to be given serious consideration.

A decade of experimental work has shown that experimentally induced motivational states influence the content of imaginative thought in ways that can be reliably coded to yield measures of important social motives—n Achievement, n Affiliation, and n Power. The use of these methods of content analysis in studies to assess individual differences in the strengths of these motives has produced an accumulation of factual information concerning the nature of motivational influences on behavior—at least within the population of high school and college students who have been the chief subjects of study. While results on hand attest to the validity of the approach, considerable further experimental refinement of the present measures is, without question, required. Nevertheless, several explorations using these measures in different cultural settings, i.e., in situations outside of the college laboratory (Child, Storm, & Veroff, 1958; Douvan, 1958; McClelland & Friedman, 1952; Rosen, 1956), have already suggested their promise for field research, particularly for survey studies. We realize, however, that a number of weighty methodological barriers stand in the way of using thematic apperceptive measures of motivation in a national survey, and that these barriers will only be overcome through direct study in field research.

---

[1] The investigations reported in this paper are based on data obtained within a national sample survey supported by the Joint Commission on Mental Illness and Health (Gurin, Project Director). Analysis of these data was supported by two grants from the National Institute of Mental Health, United States Public Health Service, Projects M2181 (Veroff, Principal Investigator) and M2280 (Gurin, Principal Investigator).

The questions of immediate interest in this paper, therefore, have to do with methodological problems in using the thematic apperceptive method—known to require a modicum of verbal ability and known to be sensitive to the situational differences at the time of administration and known to require skill in content analysis—under the conditions of national survey interviewing. These questions are listed below.

Can a single set of pictures be selected that are sufficiently relevant to the life experiences of people in all strata of society to provide fair and unbiased measures of motivation? Can the conditions of administration, in this case the interview situation, be sufficiently controlled to approximate the standardized condition of experimental studies? Can the scoring of the content of a heterogeneous sample of imaginative stories be accomplished with the same high scoring reliability that has been attained in small-scale studies of more homogeneous populations? Can anticipated differences in verbal ability in a national sample be taken into account in a way that will yield unbiased indices of motivation? Can we estimate the importance of the effect of different interviewers on the projective measures of motivation? Each of these questions will be considered separately in the first section of this monograph.

In addition to these methodological questions, we will also consider some of the interpretive problems that are highlighted when these motivation measures are investigated within the context of a national sample. These will be discussed in the second section of this monograph.

Certainly our initial answers to these questions cannot be taken as final. We consider this paper as no more than an opening wedge, explicitly pointed to laying a firmer groundwork for further integration of insights derived from clinical, experimental, and field studies of human motivation.

METHODOLOGICAL PROBLEMS

Selection of Pictures

The particular pictures used to assess motivation by thematic apperception have important effects on the motivational content of stories (Birney, 1958; Haber & Alpert, 1958; Jacobs, 1958). Pictures will differ, depending on their content, in the average amount of motivational imagery they elicit from any group of subjects (*S*s). The amount of motive-related imagery that a picture elicits from a person is influenced both by the strength of his motive and by his past experiences in settings like the one portrayed in the picture.[2] Variability in over-all motivation scores which can be attributed to peculiar characteristics of the pictures used rather than to the strength of motive within the *S*s is a source of error we seek to avoid.

The population in the survey study to be reported in this paper is a sample of 1,619 adults (21 or over), a cross-section of Americans living in private households, selected by means of probability sampling (Kish, 1953). To construct a battery of pictures that will have similar significance for all social groups represented in this sample was the major problem of picture selection that we faced. Men and women, old people and young people, blue collar workers and white collar workers, all have had experience in different sets of life situations, often so different that the motivational significance of particular kinds of pictures is likely to vary considerably from one of these groups to another. Indeed, the sex difference is so important a factor (Angelina, 1955; Davenport, 1953; Field, 1951; Morrison, 1954; Veroff, Wilcox, & Atkinson, 1953; Vogel, 1954) that it was necessary from the very beginning to plan a different set of pictures for men and women. We hoped, in addition, to overcome other systematic biases in pictures that might be attributable to many other differences in social background, in order to allow each person a fair opportunity to express his motivation in the test as a whole. Only then could the motivation scores of different social groups be legitimately compared; this aim was a compelling interest in the initiation of our study.

---

[2] Atkinson (1958, Ch. 42) has presented a theoretical scheme showing how cognitive expectations based on past experience in situations like the one portrayed in a particular picture together with the motives of the individual jointly determine the motivational content of an imaginative story.

How then did we go about selecting pictures for men and women separately? The list of pictures employed to study n Achievement, n Affiliation, and n Power (Atkinson, 1958, Appendix III) attests to the wealth of data about many pictures from assessment studies of this kind, largely on the college population. From this list we rejected pictures that were clearly oriented to the college setting (e.g., classroom scenes). Since we were interested in selecting pictures that would be relevant to three motivational contents (achievement, affiliation, power) at one time, we also rejected pictures suggesting themes that contained only one particular kind of motivational content. Initially our choices were guided only by intuition.

Before an actual pretest of some pictures in a doorstep interview, however, a more systematic evaluation of pictures was undertaken. It has been shown that ratings of pictures for strength of achievement-relatedness are correlated with the n Achievement scores later obtained from stories written about a picture by other groups of *S*s (Birney, 1958; Haber & Alpert, 1958; Jacobs, 1958). Consequently we employed a rating procedure to ascertain empirically the cue value of many pictures. A group of college students at the University of Michigan rated 40 pictures for the salience of achievement, affiliation, power, and seven other motivational concerns. At least eight *S*s rated each picture. From these ratings we tried to select pictures in which one kind of motivation was rather strongly suggested but at least one or the other kinds of motivation was weakly suggested.

Following this preliminary selection of pictures, we faced more directly the potential bias of picture cues for one social group compared to another. We hoped to eliminate bias in the instrument attributable to pictures that allow for adequate expression of motivation by only certain groups of people. Our decision, therefore, was to select pictures portraying separately for men and women a *variety* of life situations with which, in one way or another, most people in this country have had some direct contact, viz., different types of work situations and interpersonal situations. Sampling from this variety of what appeared to us to be fairly universally-relevant situations should minimize the bias in the measuring instrument.

Combining the ratings given by the students and our a priori conception of diverse situations for sampling pictures, we selected a large group of pictures for pretest in a Detroit residence survey. The stories told to these pictures were then scored for n Achievement, n Affiliation, and n Power. The pictures selected from this group for the final test forms, one for the men and one for the women, were those producing an adequate balance of scores for the three motives. The pictures finally decided upon are listed in order of presentation[3] (numbers refer to the list in Appendix III, Atkinson, 1958):

Male Form:

1. Two men (inventors) in a shop working at a machine. (2)
2. Four men seated at a table with coffee cups. One man is writing on a sheaf of papers. (101)
3. Man (father) and children seated at breakfast table. (102)
4. Man seated at drafting board. (28)
5. Conference group. Seven men variously grouped around a conference table. (83)
6. Woman in foreground with man standing behind and to the left. (103)

Female Form:

1. Two women standing by a table and one woman is working with test tubes.
2. Woman (mother) seated by girl reclining in chair.
3. Group of four women. One standing, the others seated facing each other.

---

[3] These forms are available (4″ × 6″ pictures). They may be obtained by writing to the Survey Research Center, Ann Arbor, Michigan. Price $1.00.

4. Woman kneeling and applying a cover to a chair.
5. Two women preparing food in the kitchen.
6. Same as 6 above.

A striking feature of these pictures is that they all provide only moderate structure to the situation portrayed. The set as a whole is much less strongly cued for one particular kind of motivation than sets of pictures used in experimental studies which have focused on only one kind of motivation. Rejected pictures were ones that were considered either under- or overstructured. The understructured pictures (e.g., a person standing alone in a doorway) did not provide sufficient cues for a story in a heterogeneous population. The overstructured pictures (e.g., a woman typing in an office) elicited stories with too little variability of response.

These two forms were administered to a group of 98 college students at the University of Michigan. The measures were administered in a latin square design controlling the order of presentation of pictures. Unlike the survey interview setting, the administration called for written stories according to procedures outlined elsewhere (Atkinson, 1958, Appendix III). We attempted to ascertain in these administrations: (*a*) whether possible influences on the total motivation scores of the order of presentation of pictures existed, and (*b*) whether the rough equivalence in sensitivity to n Achievement, n Affiliation, and n Power found in a sample survey held up under procedures normally employed on a sample of college students. This study showed that order of presentation of pictures did not have any systematic effect on total motive scores and that the frequencies with which achievement, affiliation, and power imagery appeared in response to each form were approximately equal in the college group. The results of these several pretests seemed to justify going ahead with the proposed set of pictures in an initial study of a national sample.

We realize, however, that the pictures finally selected undoubtedly still have differential cue values for some respondents. Negroes, for example, have to respond to situations depicting white people. Very old and very young people have to make up stories about figures that are obviously not so old nor so young as they are. The work scenes do not include farming, surgery, ditch digging, sales, and countless other occupations, although both blue collar and white collar work situations are included. Despite these differences, our working assumptions are that the situations portrayed in the pictures finally selected are not altogether foreign to most of the people with whom we are dealing, and that, by and large, the pictures will be as suggestive of the three motivational content areas to a person of one social stratum as to a person in another. We had finally to recognize that only through intensive study of the motivational cue value of these pictures for different social groups would we sharpen our insight and discover criteria that might enable us to construct a better set and to evaluate the over-all merits of the present set of pictures. In the second section of this monograph, this problem is raised again. In that section we present motive scores for various subgroups in the national sample. Differences we may find in the scores of various subgroups can be attributed to biased selection of pictures. Whether these differences in fact reflect something more than biases in picture selection is a problem we consider in that section.

The procedures followed in administration of the pictures and in the analysis for motivational content will be described subsequently. Table 1 presents some results from the national survey that are pertinent to an evaluation of the selection of pictures. Table 1 shows the frequency of imaginative stories told in response to each picture which contained each kind of motivational content.

Both male and female forms appear more strongly cued for affiliative imagery than for achievement or power imagery. To conclude, however, that the greater affiliative response is indicative of some trend in national character would require the gratuitous assumption that each set of six pictures represents equally the populations of

TABLE 1

PERCENTAGE OF STORIES CODED FOR ACHIEVEMENT, AFFILIATION, AND POWER IMAGERY (BY PICTURE AND SEX)

Sex	Picture	Affiliation Imagery	Achievement Imagery	Power Imagery
Men ($N=597$)	1	3%	32%	11%
	2	16	23	14
	3	17	2	38
	4	53	13	1
	5	33	13	30
	6	33	3	22
Average Raw Score[a]		4.85	2.90[b]	2.84
Standard Deviation		3.65	3.29	2.73
Women ($N=774$)	1	2%	45%	10%
	2	56	3	26
	3	44	10	6
	4	6	36	1
	5	40	30	6
	6	35	1	25
Average Raw Score[a]		5.31	3.97[b]	1.84
Standard Deviation		3.54	3.30	2.19

[a] Scores based on manuals and procedures in Atkinson (1958, Chs. 12, 13, 14).

[b] The category, "Unrelated Imagery" was scored 0 instead of −1. Scores obtained using 0 and −1 correlate .96.

life situations in which n Achievement, n Affiliation, and n Power are normally expressed. The differences among average scores can more appropriately be used as an index of the relative sensitivity of the present pictures to each of the motives and as a guide for improving the selection of pictures in subsequent studies.

In light of Haber and Alpert's (1958) definition of a strongly cued picture as one which elicits a particular kind of motivational imagery from 75% or more of the *S*s in a sample and a weakly cued picture as one which elicits imagery from between 25% and 50% of *S*s, we must conclude that the two six-picture forms employed in the present study are weakly cued for each of the three motives. None of the pictures can be considered strongly cued for any motive; and no more than four pictures in the set fall in the range Haber and Alpert designate as weakly cued for a given motive. Consequently, in the present study there should be fewer "false positives," which often occur when strongly cued pictures are used. On the other hand, the relatively low frequency of response has two important negative implications. First, scores from weakly cued pictures are necessarily less variable and therefore by themselves have less potential for generating strong relationships with other variables. Secondly, low average frequency of response to pictures is one of several considerations[4] that preclude the possibility of a meaningful split-half analysis of the reliability of the motive scores in this study. A split-half estimate of the reliability of the present measures is considered inappropriate because the average frequency of response is so low that even moderately motivated *S*s often respond to only one or two pictures of the set.

Control of the Interview Setting for the Projective Measures of Motivation

In experimental studies using projective measures of strength of motivation, the experimenter controls the amount of time the

---

[4] Earlier systematic split-half and test-retest estimates have yielded coefficients as high as .78 (McClelland, et al., 1953) and .74 when all the pictures were strongly cued for n Achievement and .53 (Haber & Alpert, 1958, p. 661) when all the pictures fell in the 25–50% (weakly cued) range for n Achievement. In these systematic studies of reliability, care was taken to construct subsets of pictures having qualitatively as well as quantitatively similar cues. In the present study the six pictures were not deliberately selected with an eye to having two equivalent subsets of three pictures with similar content for each motive. Rather, the choice of pictures was guided by just the opposite orientation of (*a*) presenting a variety of nonoverlapping life situations in which (*b*) three different motives might be expressed. For a fuller discussion of reliability of motivation scores, as distinct from reliability of coding which is taken up at length in a subsequent section, see Atkinson (1958), Haber and Alpert (1958), McClelland (1958), and Reitman and Atkinson (1958).

S spends telling a story. When written forms of the test are employed, each *S* is given a booklet containing a sheet of paper for each story. On each page are four sets of questions, adapted from Murray (1953), which help to remind *S* to cover all elements of a plot:

> What is happening? Who are the persons?
>
> What had led up to this situation? That is, what has happened in the past?
>
> What is being thought? What is wanted? By whom?
>
> What will happen? What will be done?

Recognizing that we could not ask a great many of the potential respondents in a national survey to write out the story, we were confronted with the task of fitting a procedure that was as similar as possible to the one above into the *interview* setting. We had to face the fact that we would be dealing with respondents who differed greatly in verbal fluency, with an individual administration rather than the usual group administration,[5] and with interviewers having different tendencies for probing. How could we command control over these differences to provide measures of motivation that would be comparable to those obtained in well-controlled experimental settings?

In the pretest already described, we explored different wordings of the leading questions and changed them slightly to elicit the most adequate imaginative protocols. The final wording of the questions for each picture and story was:

> Who are these people? (Who is this person?) What are they (is he/she) doing?
>
> What has led up to this—what went on before? What do they (does he/she) want—how do they (does he/she) feel?
>
> What will happen? How will it end?

The interviewers were instructed to write in verbatim transcript the story told to each picture as responses to each of the questions. The series of questions for each story appeared on one page with a space after each for the answers given. Each interviewer had a set of 4″ × 6″ pictures which he was to show to the *S*. The following excerpts from the detailed instructions given to the interviewers[6] reveal our decisions (*a*) not to make this a completely spontaneous story-telling situation but to have the interviewer ask the plot-guiding questions, and (*b*) not to allow the interviewers to engage in the spontaneous probing they usually engage in with open-ended questions but rather to be very specific about the conditions under which probing was permissible and what these probes should be:

> Instructions to Interviewers
>
> In response to direct questions about himself, a person often cannot give you an accurate picture of the things he is most concerned about. It has been found, however, that for some purposes a more accurate picture can be obtained from the kinds of stories he makes up in response to pictures. This story-telling procedure, then, is an attempt to do just that. In order to be able to compare people on the kinds of stories they tell, it is very important that each interviewer follow the same procedure. That is why we are giving you the explicit and detailed instructions that follow:
>
> 1. Read the instructions for the story-telling procedure to the respondent exactly as they are written on the questionnaire.
>
> 2. Show the *R* the picture and give him *about 20 seconds* to look at the picture before you ask him the questions that go with each picture. Keep the picture in front of him as he tells the story.[7]
>
> 3. There is one page for each story to be told. The same questions appear on each page. Ask the questions *in order* for each story. *Keep to the wording of the questions, and ask all of them even if it may seem repetitious in some instances.*
>
> 4. Problems that you may encounter and what to do about them:
>
> a. If the person rejects the whole task right off the bat by saying "I can't do this" or "I just have no imagination," resort to something like: "Well, just try to make up anything you want. Remember there are no right or wrong answers."
>
> b. If, in reply to one of the questions in the story form, *R* says "I've already told you that,"

---

[5] Lindzey and Heinemann (1955) have concluded from research studies of individual and group administered TATs that they are reasonably equivalent.

[6] All interviewers were those regularly employed members of the Field Staff at the Survey Research Center. In addition to receiving written instructions, two-thirds of the interviewers received direct communications about problems arising in this interviewing procedure at regional conferences set up around the country.

[7] This is a departure from laboratory procedure in which a picture is shown for only 20 seconds before a story is told. This alteration was introduced to maintain rapport in an interview study.

you can say, "Can you tell me some more about that?"

c. When a respondent says "I don't know" in reply to any of the questions about the pictures, again you might say, "You can make up anything you want. There are no right or wrong answers."

NOTE: We do not need detailed stories. All we are interested in are responses which answer the four questions for each story. If *R* answers the four questions for each story. If *R* answers the question, even if only in a couple of words, *do not probe*. Probing should come only in the cases mentioned above (a, b, c).

We would like to hold the story-telling for all six pictures down to a time limit of 20 to 22 minutes.[8] Most people will give you brief stories, but if a person begins to give a lot of detail, remind him before the next story that he doesn't have to give long stories.

The collection of thematic apperceptive stories was accomplished in the context of a study conducted for the Joint Commission on Mental Illness and Health (Gurin, Veroff, & Feld, 1960). The larger interview dealt with personal feelings of adjustment and ways people manage to handle their adjustment problems. Previous research has conclusively shown that the content of imaginative stories is highly susceptible to conditions immediately preceding the story-telling procedure. Indeed, the experimental validity of the measures depends upon their sensitivity to situationally-induced motivation. Therefore, to approximate a "neutral" test condition,[9] the projective measure appeared at the beginning of this personal interview. It was preceded only by a few questions about leisure-time activities, used to establish rapport and to avoid what might otherwise be too abrupt a beginning if the imaginative test was given first. Another survey study (Morgan, Snider, & Sobel, 1958) which included exactly the same measure of motivation but at the end of the interview (because of special rapport problems in that interview situation) found that the stories were considerably influenced by the specific nature of the material discussed in the questionnaire before the thematic apperceptive procedure was introduced.

The rationale given to respondents for introducing the novel story-telling procedure in a survey about "modern living" was:

> Another thing we want to find out is what people think of situations that may come up in life. I'm going to show you some pictures of these situations and ask you to think of stories to go with them. The situations won't be clearly one thing or another—so feel free to think of any story you want to.
>
> For example, here is the first picture. I'd like you to spend a few moments thinking of a story to go with it. To get at the story you're thinking of I'll ask you questions like: Who are these people? What do they want? and so on. Just answer with anything that comes to mind. There are no right or wrong answers.

The interviewers reported that no major difficulty arose with this testing procedure. Many of them commented that the respondents did not appear disturbed by this novel interviewing procedure and, in fact, seemed to enjoy it. The major question we must ask about the effectiveness of the procedure, however, is whether enough standardization of interviewing was established to rule out possible interviewer effects on the measures of motivation obtained from the stories. Would different interviewers testing the same population elicit stories that were similar in motivational content?[10] This problem is treated in a later section.

Coding the Stories

Training the Coders. One of the most important methodological contributions of previous work on these projective measures of motivation has been the continued concern of research workers with attaining high coding reliability. Feld and Smith (1958) have evaluated the training procedures used to teach novices the methods of content analysis for the three motives and report that conscientious use of the training materials they describe enables coders to score stories for n Achievement, n Affiliation, and n Power with reliabilities acceptable for research purposes. They review interjudge reliabilities reported in 14 published studies using the method of content analysis to be employed in this analysis and find scoring reliabilities ranging from .66 to .96 with a median of .89. They also report interjudge reliabilities ranging from .73 to .92 with a median of .87 for 12 novice coders who had just finished learning how to score a set of familiar pictures using their training procedures (i.e., after about 12 hours of independent practice). This survey of the degree of objectivity in coding obtained in small-scale experimental studies provides a standard against which to evaluate the coding reliabilities attained in this first attempt with a national sample.

---

[8] In experimental studies of students, 4 minutes are normally allowed for each story.

[9] A "neutral" test condition is one in which the experimenter makes no deliberate attempt to arouse or relax a particular kind of motivation before administration of the test.

[10] Birney (1956) has reported evidence of an "experimenter effect" on the level of achievement-related content in stories written by college students.

A team of nine coders (three for each motive) with little or no previous contact with this kind of research was trained by the method prescribed by Feld and Smith. In addition to this training procedure, scoring seminars were held, mainly to establish consensus about scoring stories derived from pictures never before used in any study. Often the stories elicited from a new picture demand specific scoring conventions not treated in scoring manuals. Feld and Smith (1958) report that coders show a small but nevertheless statistically significant decrease in reliability when coding novel material. Novel types of imagery were thoroughly discussed in the light of previous practice and the coders annotated their scoring manuals accordingly.

To insure good coding reliabilities even further, each coder was trained as an "expert" on two pictures from the male form and two from the female form, and subsequently coded only those pictures. In this way the scoring of each individual coder contributed to potentially one-third of a particular motive score for each *S*. As a result, we ruled out bias of total scores attributable to the scoring idiosyncracies of a particular coder. Further experience was given each coder on pictures for which he was finally responsible in the course of scoring 400 protocols gathered in another study at the Survey Research Center (Morgan, et al., 1958). Check-coding was done on a sample of 240 stories from 40 *S*s by two other coders[11] who had previously been checked out as "expert" in motive scoring.

Scoring reliabilities from this "warm up" check-coding, and for others reported in this paper were computed in two ways. First, we estimated the degree of agreement in noting the presence of any achievement, affiliation, or power imagery in the story—the single most important scoring decision for the coder. This agreement is expressed as a percentage (twice the number of agreements on the presence of motivational imagery divided by the sum of the total number of times motivational imagery was coded by the two coders). The second estimate of the coding reliability was the rank-order correlation between motivation scores based on all six stories coded first by the three coders and then by the check-coding experts. Although the individual percentages of agreement on "imagery" (ranging from 50% to 90%) were not consistently as high as the customary agreements found in experimental research, we felt that the total score reliabilities (.74, .76, .87) showed that these coders had attained a degree of proficiency which, augmented by further clarification of scoring problems, was sufficient to begin systematic scoring of the stories in the national sample.

Assessing Coding Reliability of Motivation Scoring in the National Survey Sample. The nine coders, each a specialist on four pictures (two from the male form and two from the female form) for one of the three motives, coded two stories from each of the 1,619 respondents' protocols (or 3,238 stories). The criteria for scoring or not scoring a particular response are those specified in the available manuals for n Achievement, n Affiliation, and n Power (Atkinson, 1958; McClelland, et al., 1953) plus specific conventions adopted for novel pictures.

Each coder's total scoring task was divided into three nearly equal sections (about 1,080 stories to be scored within each section). Furthermore, each coder scored the protocols in a different order, never scoring two stories from the same *S* in sequence.[12] These three sections each represents a coding period for which check-coding procedures were done. Within each third of the coding, a set of protocols were designated as reliability checks. The particular stories that were to be coded by the "expert coder" were unknown to the coders. From these protocols we estimated the reliabilities, again both as percentage of agreement in the coding of the presence of motivational imagery in the stories, and as rank-order correlations of total scores obtained from six stories. After each check-coding, differences between coder and "expert" were discussed. The three reliability checks provided an opportunity to follow the progress of coding reliability over the period of several months which it took each student coder, working part-time on an hourly basis, to complete approximately 100 hours of coding. The resulting percentages of agreement and total score reliabilities in the three different periods of coding are presented in Table 2.

---

[11] Two of the authors—Sheila Feld and Joseph Veroff.

The reliabilities of the coders were remarkably consistent over the three sections of coding. There are occasional shifts upward and downward, but there are no consistent trends indicating any drastic effects of coding over a prolonged period. It is our opinion that the downward trends in reliability often seemed to reflect the "expert" coder's lack of adaptation to these particular pictures. During the course of coding, certain specific conventions were adopted to establish scoring criteria for these novel

TABLE 2

SCORING RELIABILITIES FOR THE NINE CODERS AND THE THREE MOTIVATION SCORES IN THE NATIONAL SAMPLE SURVEY

Motive and Coders	Pictures	Percentage of Imagery Agreement (60 Stories)			Total Score Reliabilities (30 *S*s)		
		Check 1	Check 2	Check 3	Check 1	Check 2	Check 3
n Achievement					.89[b]	.77	.81
Coder 1	1, 6[a]	80	82	90			
Coder 2	3, 4	71	75	63			
Coder 3	2, 5	87	55	82			
n Affiliation					.91	.76	.76
Coder 1	4, 5	80	80	89			
Coder 2	2, 6	83	89	78			
Coder 3	1, 3	80	73	80			
n Power					[c]	.72	.73
Coder 1	1, 5	[c]	75	67			
Coder 2	3, 6	[c]	73	77			
Coder 3	2, 4	[c]	77	84			

[a] To be read, Picture 1 and Picture 6 on male form and on female form.

[b] Spearman rank-order correlations.

[c] These figures were similar to those attained in the second and third reliability checks for n Power coding. Exact figures are unavailable because the check coding sheets were misplaced and cannot be re-estimated since the coders changed their original codings in the direction of the consensus after discussion.

---

[12] For a given motive each *S*'s stories were scored in such a way that two were scored by one coder during the first third of coding, two in the next third by the next coder, and two in the last by the third coder. In this way any systematic bias of coding attributable to time of coding would be randomized over the *S*s. The exception to this procedure is the set of stories in the reliability checks. All of the stories of these *S*s were coded during the same period of coding.

materials. The coders were more adapted to these changes than the expert coders, who, scoring only a sample of protocols from time to time, tended to be somewhat less consistent in decisions about "novel" material.

The extent to which this is true can be evaluated in light of a second set of reliability coefficients that were computed *after* the coder and the expert check-coder had discussed every instance of disagreement and had come to a consensus of opinion about final coding. Correlations between the coders' original scorings and this consensus of opinion were uniformly higher indicating that at times it was agreed the "expert" had been in error. The post-discussion score reliabilities for first, second, and third checks were: n Achievement —.91, .94, .88; n Affiliation—.94, .84, .81; and n Power—.–,[13] .77, .77.

Consequently, we view the reliability coefficients reported in Table 2 as conservative estimates of the objectivity of the method of analysis employed. The median percentage agreement in coding motivation imagery is 80%, about 10% less than usually attained in small-scale studies with familiar materials. All of the total score reliabilities (median = .77) fall within the range found in published experimental studies, but only two of the eight are above the median of coding reliabilities reported in smaller-scale studies (viz., .89). Considering the anticipated loss in coding reliability attributable to the use of several new pictures here for the first time, we were encouraged by these results. Nevertheless we anticipate some loss in sensitivity of the measuring instruments due to some coding unreliability.

It may be mentioned in passing that the coders' awareness that there would be intermittent checks seemed to be very important in guaranteeing even coding from them. Furthermore, it gave us an opportunity to focus attention on questionable scoring criteria. As a consequence, the coders were highly aware of certain idiosyncratic coding conventions they had adopted—some of which they eventually had to change—but some of which helped us get a better picture of how the scoring criteria apply to the stories told by this kind of broad population sample in an interview setting. These "new" conventions have been integrated into scoring manuals for this set of pictures.[14]

Correcting Motivation Scores for Variations in Length of Protocol

The correlation between motivation scores and length of protocol has been found negligible in studies with college students when 4 minutes are allowed for writing stories (Atkinson, 1950, p. 27). At least four studies, however, have shown significant relationships, requiring some sort of correction when somewhat different test conditions are employed. Ricciuti (1954) and Ricciuti and Sadacca (1955) found correlations ranging from .48 to .59 between n Achievement scores and number of words in the brief protocols of high school students when no structuring questions were asked and when only 2.5 minutes were allowed for writing stories. Walker and Atkinson (1958) found a correlation of .41 between length of protocol and fear-related motivation scores in a very heterogeneous sample of soldiers when there was less than adequate control over the actual time spent writing stories. Child, Storm, and Veroff (1958) found it necessary to correct achievement scores obtained from folk tales to remove the difference attributable to length of story. We had reason to anticipate that in this heterogeneous sample there would be gross differences in the length and adequacy of protocols that might be attributed to gross differences in verbal ability of *S*s and/or to uncontrolled differences in the behavior of interviewers. Two separate problems were anticipated: (*a*) how to eliminate protocols because of inadequate responses, and (*b*) how to correct for anticipated correlations between motivation scores and length of stories in protocols otherwise judged to be adequate for further analysis.

---

[13] First reliability estimate on n Power coding unavailable (cf. Footnote c, Table 2).

[14] Available on request from Survey Research Center (Attention: Dr. Joseph Veroff), University of Michigan, Ann Arbor, Michigan.

Elimination of Protocols. At the outset of scoring the protocols it was clear that some *S*s gave responses which had to be considered inadequate for arriving at valid motivation scores. A few *S*s tended to reject the whole undertaking, some refused to give stories to some of the pictures, some were unable or unwilling to give answers to certain questions within a story. In order to deal with this problem, the coders were instructed to keep a detailed record of "inadequate responses," i.e., not answering certain questions in a story, saying "I don't know," or "It's hard to tell," or the like. For any *S* there are 24 possibilities for an inadequate response—four questions on each of six stories. Table 3 shows the frequency distribution of inadequate responses in terms of the criterion for elimination of protocols that was employed.

What does an inadequate response mean for assessment of motivation? It means that the person gave no imaginative content that could possibly be scored for motivational content. The person, for whatever reason, has prevented the measuring instrument from being applied. If a person rejects the whole task (24 inadequate responses) his protocols are obviously not suitable for scoring. Even if the person rejects the task for only one picture (i.e., four inadequate responses in only one story), his protocols should not be considered, for he has not taken the complete test. Had he responded to that particular picture, his story might have been saturated with a particular kind of motivational imagery. But what of the other possibilities? Neglecting to answer one question in a few stories, neglecting to answer two questions in one story, and so on? In Table 3 we have grouped the various possibilities of inadequate responses according to whether they are judged to be omissions of minor signficance, possibly serious, or serious. The judgment of the seriousness of omission was made on the basis of the extent to which each kind of omission might *potentially* affect the total motivation score. We considered this question: Had the respondent not made such omissions on his protocols, could his potential total score have been appreciably higher? Not giving a response to one question in a story does not subtract appreciably from a person's potential total score for that motive. Four such omissions might make an appreciable difference, and were therefore classified as possibly serious.

In light of our experience with practices informally employed in experimental studies we eliminated *S*s judged to have possibly serious or serious omissions. Thus an *S*'s protocols were judged to be adequate when at least half the stories were complete and no more than one question was omitted from any or all of the remaining stories.

TABLE 3

Frequency of Inadequate Responses to One or More of the Four Questions Asked to Guide the Story Told about Each of Six Pictures

Adequacy of Responses	Men ($N=715$)	Women ($N=904$)
All stories complete	491	654
Omissions Judged to Be of Minor Significance		
One story has 1 inadequate response	73	89
Two stories have 1 inadequate response	23	23
Three stories have 1 inadequate response	10	8
Omissions Judged to Be Possibly Serious		
More than 3 stories have 1 inadequate response	6	7
One story has 2 inadequate responses	35	32
Omissions Judged to Be Serious		
Two or more stories have 2 inadequate responses	14	15
One or more stories has 3 inadequate responses	25	39
One or more (but not all) stories have 4 inadequate responses	20	26
All 6 stories have 4 inadequate responses	18	11

This criterion led to the elimination of 17% of the men and 14% of the women. Two things should be kept in mind in appraising the significance of this number of rejected protocols. First, we deliberately adopted a fairly stringent criterion for elimination. Secondly, these *S*s were eliminated not on the basis of just one set of responses to one picture, but on the basis of inadequacy of response *on any one of six pictures.* Consequently, we judge that the frequency of inadequate protocols is not out of line with the frequency of inadequate responses that would be obtained from any complex open-ended questionnaire in a nationwide survey setting if we were to eliminate *S*s giving an inadequate response on any *one* of a set of six complex open-ended questions.

One very important question to ask about this elimination procedure is whether it has eliminated *S*s selectively from particular social groups. The psychological variable expected to have the most important bearing on ability to answer the questions about pictures is verbal fluency. The demographic characteristic most related to this variable is education of the respondent. Table 4 shows the relationship between level of education and percentage loss of data due to inadequate protocols. The loss of data is most severe in the least educated groups. Of those who have little or no schooling, nearly one-third of the sample of men and women are rejected from further analysis. What these data suggest, therefore, is that the measuring device as here employed in a survey setting is less appropriate for certain segments of the population—particularly among the least educated groups. This is hardly a startling result. Practically any verbal measure runs into the same difficulty. Inadequate responses to open-ended attitude questions tend to stem largely from uneducated groups; inadequate answers (response sets) on certain forced choice questionnaires also are most apparent in uneducated groups (Christie, 1958).

TABLE 4

THE RELATIONSHIP OF EDUCATION TO PERCENTAGE LOSS OF SUBJECTS DUE TO INADEQUATE PROTOCOLS

Educational Level	Men		Women	
	N in Sample[a]	Percentage Loss	*N* in Sample[a]	Percentage Loss
No schooling; grade school (1–6 years)	107	32	125	28
Grade school (7–8 years)	138	25	160	18
High school (9–11 years)	139	12	209	16
Completed high school	169	12	266	7
Attended college	157	7	136	7

[a] Does not include *S*s whose educational level was not ascertained.

In the future use of these measures, therefore, one has to realize that a survey population responding to them appropriately will be biased. Any variables associated with education will have biased representation. Further examination of our data also indicates that the loss of *S*s is heaviest from the low-income groups, the occupation groups of lower status (unskilled workers in particular), and the older age levels. All of these findings are consistent with the major finding on education, since these other groups that suffer the largest depletion in *S*s are groups that have a heavy proportion of respondents with little education.

We suspect that a failure to produce a meaningful protocol may also reflect deficiencies in the kinds of social motivation that are being assessed in this study—particularly achievement motivation. Hence, we may be overestimating the average level of motivation in any group from which a great many *S*s are removed because of their inadequate protocols. This additional possibility argues against using subsequent data from this analysis to estimate *absolute* norms in specific groups, but it does *not* vitiate the possibility of comparing social groups on average motivation scores, once we recognize that we may have biased estimates in certain groups.

Correction for Correlation between Length of Stories in Adequate Protocols

TABLE 5

NUMBER OF WORDS IN SIX STORIES FOR PROTOCOLS JUDGED ADEQUATE

Number of words	Men ($N=597$)	Women ($N=774$)
80–169	32	40
170–229	117	171
230–289	147	215
290–349	127	157
350–409	77	101
410–469	45	49
470–529	23	21
530–589	11	10
590–649	11	7
650 and above	7	3
Median	291	279

and Motivation Scores. The number of words in protocols accepted for 597 men and 774 women ranged from 80 to 823 as shown in Table 5. The median number of words was 291 for men and 279 for women, or about 46–48 words per story. This is exactly the same length obtained by Ricciuti and Sadacca (1955) from 2.5-minute stories written by high school students, but is approximately half the length usually obtained in studies with college students when 4 minutes are allotted for writing each story. Although there is no way to compare directly the time needed to write and to tell stories, one may note that the approximate time allotted for each story in the instructions given interviewers (3 to 3.5 minutes) is less than the 4 minutes normally allotted in studies of college students.

The raw n Achievement, n Affiliation, and n Power scores[15] obtained from the content analysis were correlated with length of protocol separately for men and women. The results are shown in Table 6. The results show the anticipated significant relationships between length of protocol and raw motive scores, product-moment correlations ranging between .21 and .28. Clearly some correction of the raw scores is necessary to eliminate the obvious fact that a person who tells longer stories stands a greater chance of obtaining a high score based on a frequency count of particular kinds of motivational content in what he has said.

Before applying correction procedures, we simplified our considerations of the motivation scores by putting the n Achievement, n Affiliation, and n Power scores on the same scale within each sex. There is no legitimate basis for making direct comparisons of raw scores between men and women on a particular motive, or between the raw motivation scores within each sex. The varying means and standard deviations of raw scores, for each sex and on each motive, are partly a function of the pictures cues, as already pointed out. Putting each motive distribution on the same scale with the same means and standard deviations allows for certain comparisons. This was done by assigning percentile ranks to each motivation score (for men and women separately). When a number of persons had the same score, the median percentile rank of the interval covered was assigned to all tied scores.

Each distribution of percentile ranks was normalized. Percentile ranks were converted to normal deviate scores based on a mean of 50 and a standard deviation of 10 (T scores). As before each sex was considered separately.

These T score conversions of percentile ranks effectively remove the differences between average motivation scores for both

TABLE 6

PRODUCT-MOMENT CORRELATIONS BETWEEN RAW MOTIVATION SCORES AND LENGTH OF PROTOCOL

Motive	Men ($N=597$)	Women ($N=774$)
n Achievement	.28	.25
n Affiliation	.20	.24
n Power	.21	.21

Note.—All p values $<.0001$.

---

[15] A total score is obtained by summing the frequencies of particular kinds of motivational content appearing in each of the six stories (see Atkinson, 1958, Part II).

men and women,[16] but they do not correct for the influence of length of protocol. The correlations between length of protocol and T scores are all within ±.01 of the original correlations between raw scores and length of protocol.

The question as to how to remove the effect of length of protocol was resolved in light of the following argument. The need for some kind of correction is obvious. A person's chance of getting a higher score on any motive is greater the longer the story he tells; the greater the amount of imagery he produces in an imaginative story, the greater the chance of it being coded as evidence of some motive. Theoretically, the best possible correction would be one based on the average of correlations between scores for many kinds of motives and the length of protocol. This average correlation would indicate the systematic relationship between length of protocol and score on any motive. Some types of motivation, e.g., n Achievement, might be meaningful determinants of the length of a story. A person highly motivated to achieve might want to try to tell a long and complete story in order to tell the best possible story. The correlation between n Achievement (content) and length of protocol might then be expected to be higher than the correlation between some other types of motivation and length of protocol. If we were to use the correlation between n Achievement scores and length of protocol to correct scores for n Achievement, it might obscure a real relationship between strength of achievement motive and length of protocol. But the average of all possible correlations would reflect the extent to which longer stories provide greater opportunity for motivational imagery of any sort to appear.

In this study, the stories have thus far been coded for only three motives. Hence the average of the three correlations provides the nearest estimate that can be made of the extent to which length of story influences opportunity to get a high score. The average of the three correlations is .22 for men and .24 for women. In each case, this average correlation was used to determine a regression line of motivation scores on length of protocol.[17] The average "expected" motivation score was determined for each length-of-protocol interval. The average expected score at each interval was then subtracted from the obtained average score for the whole sample on each motive separately. This difference score (obtained for each motive separately) was then used as a constant correction for all individual scores within particular intervals of length. Thus, a correction factor was added to scores that were too low because protocols were short, and a correction factor was subtracted from scores that were too high because of very lengthy protocols.

The effect of this correction was to remove the systematic expected difference in motivation scores as a function of the length of protocol. In Table 7 are presented the average T scores for n Achievement in men as a function of length of protocol before and after the correction described above was accomplished. We see in this table that the corrections remove the relationship between mean n Achievement scores and length of protocol. Similar effects are obtained with women on achievement motivation and with both men and women on the other motives. The over-all effects of the corrections are summarized in Table 8 which reports the existing correlations be-

---

[16] There were significant differences among the raw n Achievement, n Affiliation, and n Power scores both within and between sexes. These are readily explained in terms of the slight differences in degree to which the sets of pictures "suggested" each of the three motivational concerns. It seemed desirable to convert these scores immediately to the same scale, i.e., percentile rank, to avoid possible misinterpretation of the meaning of different average raw scores.

[17] We suspect that these three motives—n Achievement, n Affiliation, n Power—all might be expected to be more highly related to length of protocol given the interpersonal nature of the interview, than some other kinds of motivation (e.g., hunger). Hence we are well aware of the possibility that the correction employed may "over-correct" the scores. This seemed the lesser of two evils given the compelling a priori case that a longer protocol should increase the chance of getting a high score on any type of motivational content.

TABLE 7

Example of Correction Applied to Uncorrected n Achievement *T* Scores to Remove the Correlation between These Scores and Length of Protocol
(Males Only)

Number of Words in Protocol	*N*	Mean Uncorrected n Achievement Score[a]	Correction[b]	Mean Corrected n Achievement Score
80–169	32	45.75	3	48.75
170–229	117	47.95	2	49.95
230–289	147	50.04	1	51.04
290–349	127	49.96	0	49.96
350–409	77	52.90	−1	51.90
410–469	45	53.58	−2	51.58
470–529	23	52.96	−4	48.96
530–589	11	50.09	−5	45.09
590–649	11	55.82	−6	49.82
650 and above	7	56.43	−7	49.43

[a] Scores represent normal deviate equivalents using the *T* score conversion of percentile ranks, the mean is 50 and standard deviation is 10.

[b] Correction factor obtained from the regression line based on the average of correlations between n Achievement, n Affiliation, and n Power scores and length of protocol. Correction factor equals mean n Achievement score of the whole sample minus mean "expected" n Achievement score for protocols of particular length.

tween length of protocol and the final corrected motivation scores. These correlations are all low and insignificant. The final corrected scores, no longer biased by length of protocol, will henceforth be referred to as the n Achievement, n Affiliation, and n Power scores in all further analyses.

What effect will this correction of motivation scores for length of protocol have on subsequent analyses? The relationship of motivation scores to any variable which is in turn highly related to verbal fluency, should now be uncontaminated by the factor of verbal fluency. Where previously there might have been an artifactual relationship between a given variable and motivation score (because both variables were related to verbal fluency), there should no longer be any relationship. Where a relationship between motivation score and a variable might previously have been obscured artifactually, there might now appear a significant relationship.

TABLE 8

Product-Moment Correlations between Corrected Motivation *T* Scores and Length of Protocols

Motive	Men ($N=597$)	Women ($N=774$)
n Achievement	.05	.02
n Affiliation	.00	.00
n Power	−.01	−.04

Note.—All correlations are insignificant.

To illustrate some of the potential effects of removing the influence of length of protocol, we have examined the relationship between education and the three motivation scores before and after the correction was made. We again selected educational status as a crucial example, because it should represent that variable which is most likely to reflect differences in verbal fluency. Respondents were classified according to education as follows: grade school education only or less, some high school education but no college, and some college education.

When we first relate education to length of protocol (Table 9) we find, as expected, that there are substantial differences between the educational groupings in average length of protocols. Among men, college-educated *S*s produce longer protocols than the two other educational groups; and among women, the college group is higher

TABLE 9

MEAN LENGTH OF PROTOCOL (NUMBER OF WORDS) ACCORDING TO THE LEVEL OF EDUCATION OF SUBJECTS

Education	Men ($N=593$)[a]	Women ($N=771$)[a]
Grade school	256.4	240.8
High school	276.2	264.8[b]
College	303.8[c]	290.0[d]

[a] Does not include *Ss* whose educational level was not ascertained.

[b] High school mean higher than grade school mean ($p < .05$; 1-tailed test).

[c] College mean higher than high school or grade school mean ($p < .05$; 1-tailed test).

[d] College mean higher than high school mean ($p < .01$; 1-tailed test).

than the high school group which in turn is higher than the grade school group.

From these results we are led to anticipate that motivation scores will be related to education differently before and after correction. Table 10 presents the mean n Achievement, n Affiliation, and n Power scores for men in the three education groups before and after correction for differences in length of protocol. Table 11 presents a similar comparison for females.

The correction does not have any sharp differential effect on the majority of relationships between motivation scores and education before and after correction. In two instances, however, the correction does make a difference. Among men, there is no significant difference between the n Achievement scores of high school and college groups after correction where there was a significant difference before correction. Among women, the significant difference between the grade school group and the other groups in n Power scores disappears after correction. The correction seems to have eliminated a possibly artifactual difference in these two instances.

The remaining significant differences in motivation score attributable to educational status are consistent before and after correction. Evidently these are relationships between motivation and education which are independent of verbal fluency. We can, at least, feel confident that the differences obtained after correction are not attributable to differences in verbal fluency.

The Effect of Interviewers on Motivation Scores

Although we made a serious effort to standardize the interviewing procedure, we realize that having a group of 159 different interviewers (27 men and 132 women) collecting the protocols must necessarily introduce some variability into the measures that has hitherto been minimal when all protocols are collected at one time in a group administration. An estimate of the extent of this interviewer variability will help us to gain some perspective concerning how far the national survey interview departs from standardized administration procedures.

To establish such an estimate in survey data is a difficult problem. A valid test of

TABLE 10

MEAN N ACHIEVEMENT, N AFFILIATION, AND N POWER *T* SCORES OF MALES BEFORE AND AFTER CORRECTION FOR VERBAL FLUENCY

Motive and Education	Mean Score[a]	
	Before Correction	After Correction
n Achievement		
Grade school	48.76	49.90
High school	49.49	50.23
College	51.95[b]	51.92[c]
n Affiliation		
Grade school	49.21	49.88
High school	49.97	50.22
College	51.27	51.27
n Power		
Grade school	50.97	51.38
High school	49.58	49.94
College	50.23	50.09

[a] Does not include *Ss* whose educational level was not ascertained. Grade school $N=176$, high school $N=271$, college $N=146$.

[b] College mean score significantly higher than high school and grade school means ($p < .05$).

[c] College mean score significantly higher than grade school mean score ($p < .05$).

TABLE 11

Mean n Achievement, n Affiliation, and n Power T Scores of Females before and after Correction for Verbal Fluency

Motive and Education	Mean Score[a]	
	Before Correction	After Correction
n Achievement		
Grade school	48.50[b]	49.46[b]
High school	50.41	51.18
College	51.49	51.67
n Affiliation		
Grade school	47.82[b]	47.86[b]
High school	50.82	50.33
College	51.57	50.43
n Power		
Grade school	48.77[b]	49.89
High school	50.50	51.09
College	51.56	51.47

[a] Does not include *Ss* whose educational level was not ascertained. Grade school $N = 226$, high school $N = 428$, college $N = 127$.

[b] Grade school mean score significantly lower than both high school and college mean score ($p < .05$).

the effect of different interviewers requires random assignment from the population of respondents to the interviewers. This is decidedly not the case with the Survey Research Center's field operation. All interviewers are assigned respondents within their general geographical locality, and sometimes more specific locations are assigned particular interviewers. Especially in a metropolitan area, there are gross differences between the sample of respondents interviewed by one interviewer contrasted with another: some interviewers specialize in low income, foreign speaking, or Negro areas. While the selection of respondents is completely determined at the central office of the Survey Research Center by probability sampling techniques, these respondents are not assigned to interviewers at random. To the degree that the assignment is not random, there may be systematic differences in responses obtained by various interviewers that stem from differences in the composition of the samples interviewed, and not from differences attributable to the interviewers themselves. In spite of this danger of finding effects that are really attributable to respondents, we wanted to arrive at some estimate of interviewer effect. We therefore designed an analysis of "possible" interviewer effects which would be least subject to criticism that we had grossly confounded interviewers and characteristics of respondents. Our analysis undoubtedly still can be criticized on this score. But in any case, a violation of the assumption that the respondents of each interviewer represent a random sample of the population should in most cases contribute to an overestimation of interviewer effects. As a result the estimates to be described can probably be viewed as *maximal* estimates of interviewer differences.

We have examined possible interviewer effects in a limited sample of the total population of interviewers. The limitations introduced to approximate the assumption of equivalence among respondents for each interviewer were:

1. Interviewers to be considered should be white females. We eliminated from analysis all Negro and male interviewers because both are likely to be assigned to very specific populations, Negro interviewers to Negro residences and male interviewers to poorer residences.

2. Interviewers to be considered should be doing their work in nonmetropolitan sampling points. The primary sampling unit (PSU) in this survey generally corresponds to counties in states. Some of these PSUs are metropolitan. As already stated, in metropolitan areas, there is a greater tendency for the interviewers to be selectively assigned to various subpopulations. We avoid the grossest errors in meeting the assumption of random assignment of respondents by not considering metropolitan PSUs.

3. Interviewers to be considered should have interviewed at least four male respondents and/or four female respondents who gave story protocols judged adequate for motivational analysis. A study of interviewers who interviewed fewer than four *Ss* would provide too unreliable an estimate of interviewer effect.

4. Interviewers to be considered should be paired with a qualifying interviewer (not eliminated by Criteria 1, 2, and 3 above) from the same PSU. We imposed this limitation to allow for interviewer differences *within* PSUs. Our failure to meet the assumptions of random selection of respondents among interviewers should be somewhat compensated by the fact that we are dealing with

a relatively homogeneous (nonmetropolitan) population of respondents in which possible interviewer effects will be examined.

Given these limitations, we were able to find 8 pairs of interviewers for male *S*s and 15 pairs of interviewers for female *S*s. Each pair represents a different PSU. We designed an analysis of variance of motivation scores broken down into the following components: variance attributable to *S*s (or individual differences), variance attributable to interviewers, and variance attributable to PSUs. The procedure[18] for the analysis is described by Anderson and Bancroft (1952). The component variance attributable to interviewers is segregated from the *S* variance and its percentage contribution to *S* variability can be estimated. Separate analyses of variance were conducted for male and female *S*s for each of the three motives. In these analyses, we estimate the significance of the variance assignable to interviewer effects and, in addition, we segregate the component variance for interviewers and examine the extent of its effect on score variability. Table 12 presents these analyses for men and Table 13 for women.

Tables 12 and 13 show that a "possible" interviewer effect on motivation scores is statistically significant in three instances: n Achievement scores for women, and n Power scores for both men and women. Furthermore, when component variance for interviewers is specifically segregated, its effect on score variability is estimated as an intraclass correlation from which we can derive the proportion of variance in motivation scores possibly contributed by interviewer differences (Snedecor, 1946, p. 243). These were negligible for men except for n Power where the percentage increase in variance was .29. Among women, the percentage increase was negligible for n Affiliation but .18 for n Achievement and n Power. Considering these estimates as maximal estimates, we feel safe in concluding that we have achieved a fair degree of

TABLE 12

ANALYSES OF VARIANCE OF MALES' MOTIVATION SCORES
(*S*s within Interviewers within PSUs)[a]

Source of Variation	*df*	*MS*	*F*	Component Variance for Interviewers	Intraclass Correlation[b]
n Achievement					
PSU	7	96.77			
Interviewer within PSU	8	97.78	1.38	6.81	.09
*S*s within interviewer	48	70.42			
Total	63				
n Affiliation					
PSU	7	186.00			
Interviewer within PSU	8	135.13	1.33	8.47	.08
*S*s within interviewer	48	101.27			
Total	63				
n Power					
PSU	7	220.00			
Interviewer within PSU	8	170.13	2.65*	26.51	.29
*S*s within interviewer	48	64.06			
Total	63				

[a] Limited to cases of white, female interviewers of nonmetropolitan PSUs who interviewed at least four *S*s giving adequate measures.

[b] Ratio of interviewer component variance to *S* variance plus interviewer component variance.

* $p < .05$.

---

[18] We are indebted to Leslie Kish for his suggestions with procedure.

TABLE 13

ANALYSES OF VARIANCE OF FEMALES' MOTIVATION SCORES
(*S*s within Interviewers within PSUs)[a]

Source of Variation	*df*	*MS*	*F*	Component Variance for Interviewers	Intraclass Correlation[b]
n Achievement					
PSU	14	118.00			
Interviewer within PSU	15	174.40	1.90*	20.70	.18
*S*s within interviewer	90	91.24			
Total	119				
n Affiliation					
PSU	14	90.14			
Interviewer within PSU	15	72.13	.88	−2.33	−.03
*S*s within interviewer	90	81.44			
Total	119				
n Power					
PSU	14	86.14			
Interviewer within PSU	15	107.27	1.88*	12.61	.18
*S*s within interviewer	90	56.84			
Total	119				

[a] Limited to cases of white, female interviewers of nonmetropolitan PSUs who interviewed at least four *S*s giving adequate measures.

[b] Ratio of interviewer component variance to *S* variance plus interviewer component variance.

* $p < .05$.

standardization in interviewing techniques for these measures of motivation. The three significant interviewer effects will be examined in later work to explore their potential psychological significance. For the present we must recognize that we have some evidence of possible influence of interviewers on some motivation scores but apparently the influence is not substantial enough to warrant a great deal of concern.

One reason these maximal estimates of interviewer effects on motivation scores are low is that the motivation scores have been corrected for length of protocol. Similar analyses of variance were run using number of words in the protocol as the dependent variable. Table 14 shows that for both men and women respondents there is a substantial interviewer effect on length of protocol. The effect is especially strong among male *S*s. The correction of motivation scores to eliminate the effect of differences in length of protocol has removed most of the possible interviewer effect on these scores.

Over-all Evaluation of Methodological Problems Encountered

These, then, are the methodological problems encountered in the survey setting and our evaluations of the extent to which they remain as problems or as sources of error in subsequent uses of the thematic apperceptive measures of motivation. One problem was solved—the bias of motivation scores attributable to verbal fluency. One problem appears very difficult to solve—the inadequate protocols produced by uneducated persons. A third problem should be resolved by further empirical analysis—possible bias in the selection of pictures. The two remaining problems—lower coding reliability than in experimental studies and possible bias in motivation scores attributable to interviewers—are ones that lower the reliability of measurement in subsequent analyses using the motivation scores. But both of these sources of error appear to be ones that certainly can be greatly reduced in subsequent studies. Experimental in-

TABLE 14

ANALYSES OF VARIANCE OF MALES' AND FEMALES' LENGTH OF PROTOCOLS (*Ss* within Interviewers within PSUs)[a]

Source of Variation	*df*	*MS*	*F*	Component Variance for Interviewers	Intraclass Correlation[b]
For Males					
PSU	7	13.91			
Interviewer within PSU	8	12.92	12.99**	2.96	.73
Ss within interviewer	48	1.06			
Total	63				
For Females					
PSU	14	9.94			
Interviewer within PSU	15	3.50	2.00*	.67	.28
Ss within interviewer	90	1.75			
Total	119				

[a] Limited to cases of white, female interviewers in nonmetropolitan PSUs who interviewed at least four *Ss* giving adequate measures.

[b] Ratio of interviewer component variance to *S* variance plus interviewer component variance.

* $p < .05$.

** $p < .001$.

vestigations can be conducted to clarify the difficult scoring questions that arise with new pictures, as with those employed here for the first time, and to reduce the error in coding that is a normal consequence of trying to fit general scoring criteria to imagery produced by novel stimuli. Similarly, the influence of interviewers, particularly on length of protocols, should be greatly reduced by even more thorough briefing of interviewers. We can be guided by the results of this study concerning the conditions of administration, and, specifically, the desired length of stories.

INTERPRETATION OF MOTIVATION SCORES

If methodological problems of the sort we have discussed in the previous section are seriously considered, the use of thematic apperceptive measures of motivation, and other sensitive techniques of personality assessment, can be expected to begin to contribute to the study of important substantive problems in survey research. Yet even after methodological precautions are taken, there remains an important question: What conceptual meaning do we attribute to the motivation scores? This is the issue now to be considered.

The problem of interpretation of motivation scores obtained from thematic apperception is not a new one (Atkinson, 1958; McClelland, et al., 1953); but the broader setting of a national survey study exaggerates the difficulties that are normally encountered even in fairly well-controlled experimental studies. Interpretive problems are particularly highlighted, as we will attempt to point up in this section, when we examine the relationships between demographic characteristics and motivation scores. The national survey which enables us to compare the motivation scores of the different subgroups of the larger population also serves to accent theoretical and interpretive problems that are often glossed over in laboratory research.

Three factors are thought to influence the frequency of motivational imagery produced in stories elicited by pictures (taken as the measure of the strength of a particular kind of motivation): (*a*) individual differences in strength of motive, a relatively general and enduring disposition within the person to strive for certain kinds of goal satisfaction, affect motivational imagery; (*b*) cues in the pictures that are used to elicit imaginative thought may also heighten or dampen motivation for particular goals to the extent that the situation portrayed is relevant to the life experience of the respondent; and, (*c*) the immediate life situation of the person may arouse or dampen motivation for particular goals. The latter influence may be some factor in the immediate test situation of the respondent, or some more general but transient influence on his life situation, or some more general and longer lasting influence on his life situation (e.g., constant achievement pressure of the job) at the time the measure is administered. Both pictures and a person's life situation provide cues that define expectations concerning what kinds of goals are relevant and reachable, and how attractive it would be to reach the goals.

In previous sections we have discussed the steps taken to minimize or control the influence of some situational and picture effects in order that motivation scores obtained in this survey could be interpreted with some confidence as manifestations of individual differences in strength of motives. Many of these steps (i.e., control of the interview setting, attention to the type of pictures employed, corrections for differences in length of protocols) do, undoubtedly, enhance the possibility of viewing the obtained scores as measures of differences in personality. Nevertheless, we believe this assumption should still be considered tentative.

While we attempted to remove biasing effects of pictures, we have no independent evidence that we have in fact removed all possible bias. For instance, can we assume that the pictures employed, which portray relatively young men and women in common life situations, will have the same meaning for young, middle aged, and aged respondents? Furthermore, the standardization of the interview situation might control the immediate situational influences at the time the test was administered but obviously could have no effect on life conditions affecting the person at the time the survey was conducted. Can we assume, in advance, that the active young businessman is being tested under situational circumstances that are comparable to those of an older man enjoying retirement from business and the stress of day-to-day activities? Or must we conceive of the possibility that these gross differences in the total life situation in which the administration of our test of motivation is embedded will produce effects comparable to those produced by experimental manipulation of situational influences in studies of motivation in college students?

Therefore when our results show that a particular kind of motivation is stronger in one social group than another, our task is not finished; rather, the difficult task of interpretation is now thrust upon us. We may, in light of earlier research, assume that the combined effect of picture cues, situational pressures, and dispositional differences is greater in one group than in another. We have located a difference in motivation which now requires some kind of theoretical interpretation. The inference that the result means the two groups differ in strength of a general and enduring disposition of personality requires the assumption, and supportive argument, that the effects of picture cues may be considered equivalent in the two groups and that the situational pressures for expression of the motive, whether transient or relatively enduring, are also equivalent. Often the arguments in support of one or another assumption will appear thoroughly convincing. For example, the assumption has generally been made in laboratory experiments with college students that in view of the relative homogeneity of both their present and past life conditions, differences in motivation scores could be assumed to reflect dispositional differences. Just as often, however, we shall find ourselves with a problem of interpretation which cannot be settled within the framework of the present results. In these cases, the identification of important motivational differences and the ambiguity in the interpretation will define an important problem for future research.

It can be argued that certain life situations can have lasting effects on a person's disposition so that "situational" differences become "personality" differences. Little is known concerning how permanent are the effects of certain life experiences, particularly ones that occur later in life. The assumption of the primacy of early experiences in the generation of personality dispositions has been widely accepted. Facing frankly the paucity of evidence concerning dispositional versus situational factors in motivation, and their interactions, we can see that any one of the following arguments might be pursued in the interpretation of obtained differences between certain social groups.

1. The obtained difference is attributable to differences in enduring personality dispositions acquired early in life.

2. The obtained difference is the consequence of a change in personality disposition induced by important situational factors later in life.

3. The obtained difference is a consequence of exposure to differential temporary situational pressures affecting the level of aroused motivation regardless of equivalence in underlying personality disposition.

Thus two problems remain that must be faced in confronting the substantive results obtained with the thematic apperception instrument: assessments of the bias of picture cues and the influence of the *general* life situation on the motivation scores. These problems will be illustrated in the discussion of substantive results that follow.

In Tables 15 to 22 are presented some of the obtained differences in n Achievement, n Affiliation, and n Power scores when certain social groups in American society are compared. The complete interpretation of these and many other results is reserved for subsequent papers. The issue of alternative interpretations can be dealt with thoroughly only when appropriate demographic controls are introduced and the whole pattern of concurrent findings is considered. These tables are presented in the present context for two reasons: first, to illustrate concretely the issues which arise when we are finally confronted with the task of arriving at a valid interpretation of the motivation scores; and secondly, to provide researchers who are interested in particular social groups some descriptive information concerning potential differences in motivation among these groups. Under no circumstances should these data be considered unequivocal norms about strength of motives in particular social groups until the issue of the extent to which obtained differences are situational, dispositional, or artifactual is settled. Furthermore, under no circumstances should all of the differences presented in these tables be considered reliably different from one another. We have highlighted some that illustrate the problem of interpreting motive scores, but we have reserved the problem of statistical significance for the substantive appraisal of results in future articles. For readers who would like some bearings on the approximate differences in percentages required for a statistically significant comparison between two independent groups, we have included in the Appendix a table of differences required for significance (.05 level) with different sample sizes (*N*s). This table is based on approximate sampling errors of differences in percentages. Some of the analyses of the differences in percentages in the following tables require statistical treatments other than what is involved in using the table in the Appendix (correlation, matched comparisons, one-tailed tests of significance). This Appendix table is presented merely to provide approximate estimates of the statistical significance of obtained differences.

In all of the tables, *S*s have been classified as "High" or "Low" in motivation score by dividing the distribution of corrected *T* scores as near to the median of the distribution for the whole sample as was possible. The tables indicate what percentage of particular subgroups are high (above the median) in the national sample. Direct cross comparisons between men and women are not justified since different test forms were employed for the two sexes and the distributions of scores for men and women were considered separately in designating high and low scores.

Education

In an earlier section we examined the correction for differences in length of protocol as it affected the relationship between level of education attained by the *S*s and mean motivation scores. In Table 15 the same result is presented in terms of the percentage of scores in each of three educational levels that are above the median for the whole sample for each motive. The percentage of college educated *S*s, whether men or women, with high n Achievement scores is substantially higher than the percentage of *S*s who have had only a grade school education. Among women, the same result appears in connection with affiliation scores and power scores.

Now how would one interpret these differences? Are the picture cues biased in favor of the college population? Has the life experience of a college education affected the achievement motive of college women? Or were college women high in achievement motivation from an early age, and thereby managed to get to college? We cannot resolve these different views. Later evidence, however, may bear on which interpretation is more appropriate.

TABLE 15

EDUCATIONAL LEVEL RELATED TO N ACHIEVEMENT, N AFFILIATION, AND N POWER

Motive[a] and Education	Percentage[b] of High Scores[c]	
	Men	Women
n Achievement		
Grade school	48	44
High school	49	49
College	62	52
n Affiliation		
Grade school	47	43
High school	47	53
College	51	52
n Power		
Grade school	55	42
High school	47	51
College	49	54

[a] Here and in all subsequent tables, the motive measures used are the *T* scores corrected for verbal fluency.

[b] Does not include *S*s whose educational level was not ascertained. Grade school *N*s, Men = 176, Women = 222; high school *N*s, Men = 271, Women = 422; college *N*s, Men = 146, Women = 127.

[c] Here and in all subsequent tables, high scores refer to the scores on each motive measure that are above the median for the total sample of men or women.

Occupation

In Table 16 are presented the percentage of men presently employed on a full-time basis in various occupations who obtained high scores for each of the three motives. Achievement motivation scores are much more frequently high among persons in higher status occupations than among men in lower status occupations.

No systematic relationship between n Affiliation or n Power scores and occupational *status* is clearly apparent, though certain possibly meaningful differences in the different occupations may be discerned. Whereas 51% of managers and proprietors score high in power motivation, only 43% of professionals and 40% of clerical workers, the remainder of the white collar occupation category, score high in power motivation. Should the difference here be considered a manifestation of differences in basic personality disposition which has led one group of men to positions of supervision and another not, or should the difference in motivation scores be considered a reflection of the difference in motivational

TABLE 16

OCCUPATIONAL LEVEL RELATED TO N ACHIEVEMENT, N AFFILIATION, AND N POWER (MEN EMPLOYED FULL-TIME ONLY)

Occupation of Respondents	*N*[a]	Percentage of High Scores		
		n Ach	n Aff	n Pow
Professionals	67	60	54	43
Managers and proprietors	70	59	56	51
Clerical workers	30	57	47	40
Sales workers	34	59	47	47
Skilled workers	120	50	55	50
Semiskilled workers	88	52	41	53
Unskilled workers	38	45	37	50
Farmers	39	44	31	51

[a] Does not include *S*s whose occupational level was not ascertained or *S*s whose occupations were not codable into these categories.

influences contained in the day-to-day life situation of managers, as a group, versus those classified as professionals? Or, could the result conceivably be attributed to some biasing in the content of the picture cues favoring one group and not another with respect to the suggestion of power-related imagery? Again, the matter of interpretation must await the extensive analysis of these and related findings that will be the substance of future papers based on this investigation.

Income

Table 17 presents the relationships between motivation scores and family income. The irregular relationship between income and n Achievement scores in men clearly suggests that other demographic factors must be controlled in an effort to clarify the result. However, oddly, in the case of women, there is a positive relationship between achievement motivation scores and family income.

Age

Perhaps the clearest illustration of the problem of interpretation of scores as effects of personality disposition versus situation effects is apparent in consideration of age trends, particularly in the achievement motivation scores of men (Table 18).

There are some notable differences between the age groups in achievement motivation (for both men and women). For example, the highest achievement motivation is found among the youngest men, and the next highest among a middle-aged

TABLE 17

FAMILY INCOME RELATED TO N ACHIEVEMENT, N AFFILIATION, AND N POWER

Motive and Family Income	N^a	Percentage of High Scores		N^a
		Men	Women	
n Achievement				
Under $1,999	66	42	42	122
$2,000–3,999	114	54	42	187
$4,000–4,999	100	47	48	125
$5,000–6,999	158	56	50	170
$7,000–9,999	89	48	57	96
$10,000 and above	60	57	52	52
n Affiliation				
Under $1,999		42	47	
$2,000–3,999		54	47	
$4,000–4,999		41	51	
$5,000–6,999		49	55	
$7,000–9,999		52	50	
$10,000 and above		48	52	
n Power				
Under $1,999		58	43	
$2,000–3,999		49	46	
$4,000–4,999		47	48	
$5,000–6,999		48	56	
$7,000–9,999		47	57	
$10,000 and above		55	40	

[a] Does not include *Ss* whose family income was not ascertained.

TABLE 18

AGE RELATED TO N ACHIEVEMENT, N AFFILIATION, AND N POWER

Motive and Age	N^a	Percentage of High Scores		N^a
		Men	Women	
n Achievement				
21–24	32	66	47	53
25–34	165	50	50	205
35–44	119	57	50	184
45–54	123	54	45	152
55–64	82	42	52	98
65 plus	76	47	35	77
n Affiliation				
21–24		47	53	
25–34		48	53	
35–44		44	55	
45–54		49	53	
55–64		53	36	
65 plus		49	38	
n Power				
21–24		44	55	
25–34		45	52	
35–44		52	47	
45–54		51	49	
55–64		49	45	
65 plus		54	47	

[a] Does not include *Ss* whose age was not ascertained.

group (35–44 years), while the lowest achievement motivation scores appear in the older groups. A number of alternative interpretations of these differences can be offered. On the one hand, these differences may reflect some important generational differences in the strength of achievement motive dispositions. On the other hand, the stage of a man's life cycle may have a great deal to do with his current aroused achievement motivation—the eagerness of youth starting out in a career and the middle-aged pressure to "achieve or else" may both lead to heightened achievement motivation in contrast to the pressures of old age that may detract from achievement arousal. From other results we have obtained thus far, relating achievement motivation to other variables (controlling for age), we do find that the age groups vary considerably in what achievement motivation signifies. One can at the present only speculate about a life cycle or a generation difference influencing these results, and perhaps both sets of factors are crucial.

The age differences obtained may reflect picture biases as well; the possibility of systematic age differences in the connotation of the pictures cannot be excluded. Very old people, for example, might have special difficulty in generalizing from the pictures presented to their own life situations since the figures in the pictures are not clearly old in appearance.

As a result of these obtained age differences in motivation scores and our speculations about their meaning, we have considered age as a critical variable for investigations of other substantive relationships to be reported in future papers.

Race

One of the most apparent instances where subgroup differences in motivation scores may reflect the operation of picture bias is the comparison of scores for white and Negro respondents (Table 19). The pictures presented to all respondents for the most part clearly portrayed white persons. What meaning does this have for a Negro person? It has been found that the socio-economic position of the Negro can play as important a role in determination of some of his responses as does his race (Korchin, Mitchell, & Meltzoff, 1950) and that there is some tendency for Negroes to be less guarded in their responses to pictures of white persons (Cook, 1953). Other research has also explored this and related problems, but no unassailable conclusion has been drawn about the relative merits of using pictures of whites or Negroes with Negro respondents (Light, 1955; Riess, Schwartz, & Cottingham, 1950; Schwartz, Riess, & Cottingham, 1951).

Comparisons between white and Negro respondents are also limited by the immediate test situation. Some Negroes were interviewed by white interviewers, although some of them (especially those in the South) were interviewed by Negro interviewers. Previous research dictates that we make note of this difference, although its implications are not entirely clear. Schwartz, et al. (1951) find that Negroes express more ideas to a TAT if the interviewer is white.

Therefore, we do not have firm empirical ground for interpreting the meaning of the

TABLE 19

RACE RELATED TO N ACHIEVEMENT, N AFFILIATION, AND N POWER

Motive and Race	Percentage of High Scores[a]	
	Men	Women
n Achievement		
White	51	49
Negro	48	40
n Affiliation		
White	48	51
Negro	45	50
n Power		
White	48	49
Negro	61	50

[a] Does not include *Ss* whose race was not ascertained or *Ss* whose race was not codable into these categories. White *Ns*, Men = 538, Women = 688; Negro *Ns*, Men = 31, Women = 60.

differences found in Table 19. At the present we can only note that white women seem to be higher in achievement motivation scores, and Negro men are higher than white men in power motivation. The other comparisons show minimal differences. We can ask would we have obtained other differences had we used pictures of Negroes? Of course we cannot answer this question in the present study. Where we will use the scores of Negroes, we will treat them as if there were no picture biases. Later results may lead us to reconsider the problem.

Place of Residence

In Table 20 we present the distributions of motivation scores according to the place of residence—whether in a metropolitan area or their suburbs, a small city, a small town, or a rural area. These findings present some differences in scores that are not completely independent of status considerations noted in other comparisons—income, occupation, or education differences. Nevertheless there are noteworthy results: For both men and women there is greater concentration of high achievement motivation scores in small towns and rural areas than in larger communities. High affiliation motivation seems to be more prevalent among men from smaller communities, but highest in women living in metropolitan areas. Power motivation seems to be relatively high in metropolitan areas.

TABLE 20

CURRENT PLACE OF RESIDENCE RELATED TO N ACHIEVEMENT, N AFFILIATION, AND N POWER

Motive and Place of Residence[a]	*N*	Percentage of High Scores		*N*
		Men	Women	
n Achievement				
Metropolitan areas	86	48	45	104
Suburbs	87	52	42	97
Small cities	92	51	43	118
Small towns	180	56	49	152
Rural areas	225	53	49	230
n Affiliation				
Metropolitan areas		42	58	
Suburbs		44	49	
Small cities		48	49	
Small towns		48	54	
Rural areas		51	46	
n Power				
Metropolitan areas		55	57	
Suburbs		56	45	
Small cities		46	46	
Small towns		44	51	
Rural areas		50	48	

[a] Metropolitan areas and their suburbs are defined by the United States Census Bureau Classification: small cities have populations of over 50,000, small towns populations of under 50,000, and rural areas are open farm country not located in a standard metropolitan area.

These differences may have important implications for the thesis that the type of community one lives in can determine the development of certain motivational interests. For example, one might want to say that a *consequence* of living in smaller communities is the development of a heightened interest in affiliation and yet there remains one possibility that such an interpretation is in error. Motivation can play a large role in *determining* the kind of community people live in, whether they are willing to remain in a city or in a rural area, whether they migrate to suburban areas, and the like. For example, men with high affiliation motivation may prefer small towns where intimacy is easier.

Broken Home Background and Death of a Spouse

In the foregoing sections we have suggested that many of the differences in motivation scores obtained can be as plausibly interpreted as situational differences as they can as personality differences. The "correctness" of either interpretation, or the relative importance of each interpretation should be resolved in further study. There is evidence that the resolution of this problem will not be a general one; one or the other interpretation will not be uniformly appropriate, but both will be involved. This evidence comes from examining two other special groupings—broken home background, and death of a spouse.

All respondents were asked whether or not they had lived with both of their parents before the age of 16. Through further probing we were able to group respondents into three categories:

1. Coming from intact homes.
2. One or both parents died before the respondent was 16.
3. Parents were divorced or separated before respondent was 16.

These experiences in early childhood are likely to be profound catalysts for inducing basic personality patterns. We feel that differences found in adults with these three kinds of early backgrounds might reflect early personality integration of motivation. In these comparisons (Table 21), parental divorce seems to have had effect on achievement motivation—lowering it for males and raising it for females. Death in the family lowers affiliation motivation in males, but not in females, and either death or divorce raises power motivation in males and lowers it in females. We have found a sizable relationship between the age of the respondent and experiences of family disruption by divorce. Proportionally more younger than older respondents make up the group from divorce backgrounds. Even with age controlled, however, most of the differences apparent in Table 21 remain; for the most part, the differences cited occur at each age group considered.

TABLE 21

Broken Home Background related to n Achievement, n Affiliation, and n Power

Motive and Home Background	Percentage of High Scores[a]	
	Men	Women
n Achievement		
Intact home	54	48
Parent(s) died	45	42
Parents divorced or separated	38	60
n Affiliation		
Intact home	49	49
Parent(s) died	39	52
Parents divorced or separated	46	51
n Power		
Intact home	47	51
Parent(s) died	61	45
Parents divorced or separated	62	44

[a] Does not include *Ss* whose home background was not ascertained or *Ss* whose home background was not codable into these categories. Intact home *N*s, Men = 471, Women = 585; Parent(s) died *N*s, Men = 80, Women = 125; Parents divorced *N*s, Men = 24, Women = 45.

Explanations of these differences would rest, we argue, on considering what implications these early life experiences have on motive development. For example, we might suggest that divorce of one's parents can have a lasting effect on a child's achievement motivation because the disruption creates significant early experiences and attitudes towards the competence of both the father and the mother. Usually children remain with the mother when there is a divorce. Resentment or criticism of the father must be inevitably transmitted to the child. This resentment added to the mere absence of an intimate contact with the father can have opposite effects on the achievement motivation of boys and girls. A boy, having lost his masculine model for achievement, may become highly involved in avoiding failure. In so doing, his achievement motivation, his positive motivations for success, become weakened. On the other hand, girls living with a divorced mother have a readily available model for achievement identification. Resentment of the father can reinforce a need for feminine independence and self-reliance in a masculine world. The fact that her mother is apparently self-sufficient further enhances an image of the achievement orientation of women. This is overgeneralized speculation about the dynamics of divorce and achievement motivation. Nevertheless, it is an example of how we would treat differences in motives found in family background comparisons. Having obtained differences in motivation scores for these groups we can proceed with an interpretation of differences by assuming that the differences reflect enduring *personality* changes brought on by family disruptions in early childhood. What is perhaps most important is that these potential differences in motivation are attributable to dispositional differences. Therefore, the motivation score *can* reflect personality differences.

The last table we are going to consider is one which compares two groups of men and women—those presently married and those whose spouses have died (Table 22). We present this table only for older (50 and above) men and women because the number of widows and widowers under 50 is relatively low. Differences without this age control might be attributable to age rather than the loss of a spouse. There are interesting differences in the table. For all motives, except power motivation in men, the loss of a spouse yields a decrease in score which is primarily a situational change. The general conclusion might be that losing one's wife or husband in later life for most people has the effect of decreasing their motivational concerns—a conclusion not at all out of line with current thinking about problems in old age. At any rate we have evidence that a life situational factor in later life can have a strong effect on motivation scores.

SUMMARY AND CONCLUSIONS

We have reviewed some of the methodological and interpretive problems attending the use of thematic apperceptive measures of motivation in a nationwide interview study. In light of this exploration, it is possible to offer both evaluations of the difficulties encountered and a prognosis concerning these measures in substantive research within the survey setting.

Selection of Pictures. The selection of separate sets of six pictures for men and for women was guided by previous research concerning the effect of picture cues on thematic apperception. The percentage of stories elicited by these pictures which contained imagery related to achievement, affiliation, and power was lower than anticipated from pretests but sufficient, we believe, to allow for valid assessment of individual and group differences in the strength of the three motives.

We have no independent basis for evaluating the extent to which the objective was attained of providing a fair opportunity for persons in all strata of American society to express their motives. Only intensive analysis of the responses to particular pictures by particular segments of the population (reserved for a later paper) will provide the kind of evidence needed for further evaluation of the adequacy of the pictures that were employed and clarification of the issues to be faced in subsequent attempts of this sort.

Interviewer Effects. In light of the known sensitivity of thematic apperception to conditions of administration, we anticipated and found that despite our efforts to standardize the interviewing procedure, the interviewers did (apparently) contribute significantly to the variance of some (but not all) of the motivation scores. However, we were gratified to find that our estimates

TABLE 22

MARITAL STATUS RELATED TO N ACHIEVEMENT, N AFFILIATION, AND N POWER
(Age 50 and Above Only)

Motive	Percentage of High Scores			
	Men		Women	
	Married ($N=169$)	Widowed ($N=27$)	Married ($N=122$)	Widowed ($N=93$)
n Achievement	47	37	47	43
n Affiliation	50	41	43	39
n Power	48	63	47	43

of interviewer effect—maximal estimates as a consequence of possible confounding of interviewer and respondent effects in survey data—were no higher than in fact they were. That interviewer effects on motivation scores were less than anticipated can be attributed to the correction employed to remove the relationship between raw motivation scores and length of protocol. Interviewers differed considerably in the average length of the imaginative protocols they obtained. Correction of motivation scores for this factor effectively eliminated much of an otherwise substantial interviewer effect.

As a consequence of bias attributable to interviewers still remaining in some of the motivation scores, we anticipate some loss in the sensitivity of the measures (i.e., an increase in error variance) in later analyses of substantive relationships, e.g., relationships of motivation to demographic variables.

Reliability of Coding. The reliability of coding attained by a team of nine relatively novice student coders, while less than that normally attained with this method of content analysis in small-scale experimental studies of college students, was nevertheless encouragingly high. The range of coding reliability coefficients for the motivation scores assigned to individuals was .72 to .91 with a median of .77. All of these estimates of coding reliability are within the range of reliabilities reported in earlier published studies. However, in comparison with the median standard of .89 attained in experimental studies, in this study with new pictures and other novel problems there is a loss in reliability of coding and hence some increase in the error of measurement. We expect, therefore, an additional loss in the sensitivity of the measures in later analyses of substantive relationships.

Inadequate Protocols. A substantial number of individuals produced protocols that were judged inadequate for further analysis. The over-all loss of data amounted to 17% among men and 14% among women. In both sexes, the loss was heavy and serious, as might have been anticipated, only among the least educated groups. Hence, the average motivation scores of some subgroups in the population, specifically those in which level of education is very low, must be considered biased estimates—derived only from the most verbal members of those groups.

Verbal Fluency. One problem of considerable importance, the relationship between motivation scores and length of imaginative protocol, which is described by correlations of .20 to .28 in this national sample, seems to have been quite adequately overcome. The correction devised reduces these correlations to the range from −.04 to .05 and effectively removes the possibility that substantive relationships, like the significant positive relationship between n Achievement and education of respondent, can be considered spurious because of a dependency of motivation scores on verbal fluency. As mentioned earlier, this correction also produced a very substantial reduction in the extent of variability of motivation scores attributable to interviewer differences.

Interpretation of Motivation Scores. Differences in motivation scores in various subgroups of the population can be variously interpreted as reflecting differences in: personality, life situation, and the significance of the pictures for various groups. The relative importance of each of these interpretations may be evaluated only by further inquiry into the kinds of variables related to these scores within different subgroups. A particular network of interrelationships may be suggestive of which of these views may be supported. It might very well be that the interpretation of the scores will vary depending on which social group is being investigated. Or, one conclusion that seems likely is that for each social group the motivation score can be partially interpreted as personality assessment and partially as an assessment of reactions to ongoing life situations. There is evidence in the survey data that each of these interpretations may be operating in the scores. Motivation scores reflect differences both in early life experiences (broken home background) and later life experiences (death of a spouse).

These alternative considerations of the conceptual meaning of the motivation scores precludes the consideration of their relationships with demographic variables as unequivocal norms describing subgroup differences in personality. The results have been presented here with the hope that they can serve as background for exploring the significance of motivation differences in society.

If methodological precautions like those outlined here are seriously considered, the use of thematic apperceptive measures of motivation (and other sensitive techniques of personality assessment) can be expected to make marked contributions in the study of important substantive problems in survey research. Careful inquiries into the relationship of these measures to other psychological variables and to demographic indices in this survey, and in subsequent national studies, should greatly enhance our understanding of the social origins and consequences of motivation. At the same time, survey studies will provide insights that are needed to refine the thematic apperceptive method itself. The use of thematic apperception to assess motivation in survey studies of national character will mean that factual evidence concerning personality configurations within American society can be integrated with the fund of factual evidence concerning the dynamics of motivation already derived from experimental and clinical use of these same measures.

REFERENCES

Anderson, R. L., & Bancroft, T. A. *Statistical theory in research.* New York: McGraw-Hill, 1952.

Angelina, A. L. Un novo método par avaliar a motivacão humana. Unpublished doctoral dissertation, Univer. São Paulo, Brazil, 1955.

Atkinson, J. W. Studies in projective measurement of achievement motivation. Unpublished doctoral dissertation, Univer. Mich., 1950.

Atkinson, J. W. (Ed.) *Motives in fantasy, action and society: A method of assessment and study.* Princeton, N. J.: Van Nostrand, 1958.

Birney, R. C. Experimenter effect on the achievement motive. Unpublished manuscript, Amherst Coll., 1956.

Birney, R. C. Thematic content and the cue characteristics of pictures. In J. W. Atkinson (Ed.), *Motives in fantasy, action and society.* Princeton, N. J.: Van Nostrand, 1958. Pp. 630–643.

Campbell, D. T. The indirect assessment of social attitudes. *Psychol. Bull.,* 1950, **47**, 15–38.

Child, I., Storm, T., & Veroff, J. Achievement themes in folk tales related to socialization practice. In J. W. Atkinson (Ed.), *Motives in fantasy, action and society.* Princeton, N. J.: Van Nostrand, 1958. Pp. 479–492.

Christie, R. Acquiescence and education. Unpublished manuscript, Columbia Univer., 1958.

Cook, R. A. Identification and ego defensiveness in thematic apperception. *J. proj. Tech.,* 1953, **17**, 312–319.

Davenport, J. W. The projective expression of the achievement motive in relation to career orientation and social satisfaction in home economics freshmen. Unpublished master's thesis, Univer. Mass., 1953.

Douvan, Elizabeth. Social status and success strivings. In J. W. Atkinson (Ed.), *Motives in fantasy, action and society.* Princeton, N. J.: Van Nostrand, 1958. Pp. 509–517.

Douvan, Elizabeth, & Walker, A. M. The sense of effectiveness in public affairs. *Psychol. Monogr.,* 1956, **70**(22, Whole No. 429).

Feld, Sheila, & Smith, C. P. An evaluation of the method of content analysis. In J. W. Atkinson (Ed.), *Motives in fantasy, action and society.* Princeton, N. J.: Van Nostrand, 1958. Pp. 234–241.

Field, W. F. The effects of thematic apperception upon certain experimentally aroused needs. Unpublished doctoral dissertation, Univer. Md., 1951.

Gurin, G., Veroff, J., & Feld, Sheila. *Americans view their mental health: A nationwide interview survey.* New York: Basic Books, 1960.

Haber, R. N., & Alpert, R. The role of situation and picture cues in projective measurement of the achievement motive. In J. W. Atkinson (Ed.), *Motives in fantasy, action and society.* Princeton, N. J.: Van Nostrand, 1958. Pp. 644–663.

Jacobs, B. A method for investigating the cue characteristics of pictures. In J. W. Atkinson (Ed.), *Motives in fantasy, action and society.* Princeton, N. J.: Van Nostrand, 1958. Pp. 617–629.

Kish, L. Selection of the sample. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences.* New York: Dryden, 1953. Pp. 175-240.

Korchin, S. J., Mitchell, H. E., & Meltzoff, J. A critical evaluation of the Thompson Thematic Apperception Test. *J. proj. Tech.,* 1950, **14**, 445-452.

Lansing, J., & Heyns, R. Need affiliation and frequency of four types of communication. *J. abnorm. soc. Psychol.,* 1959, **58**, 365–372.

Light, B. H. A further test of the Thompson TAT rationale. *J. abnorm. soc. Psychol.,* 1955, **51**, 148–150.

Lindzey, G., & Heinemann, S. H. Thematic Apperception Test: Individual and group administration. *J. Pers.,* 1955, **24**, 34–55.

Maccoby, Eleanor, E., & Maccoby, N. The interview: A tool of social science. In G. Lindzey (Ed.), *Handbook of social psychology.* Vol. 1. Cambridge: Addison-Wesley, 1954. Pp. 449–487.

McClelland, D. C. Risk taking in children with high and low need for achievement. In J. W. Atkinson (Ed.), *Motives in fantasy, action and society.* Princeton, N. J.: Van Nostrand, 1958. Pp. 306–321.

McClelland, D. C., Atkinson, J. W., Clark, R. A., & Lowell, E. L. *The achievement motive.* New York: Appleton-Century-Crofts, 1953.

McClelland, D. C., & Friedman, G. A. A cross-cultural study of the relationship between child-training practices and achievement motivation appearing in folk tales. In G. E. Swanson, T. M. Newcomb, & E. L. Hartley (Eds.), *Readings in social psychology.* New York: Holt, 1952. Pp. 243–249.

Morgan, J. N., Snider, M., & Sobel, M. G. *Highlights from a study on lump sum redemption settlements and rehabilitation.* Ann Arbor, Mich.: Survey Research Center, 1958.

Morrison, H. W. Validity and behavioral correlates of female need for achievement. Unpublished master's thesis, Wesleyan Univer., 1954.

MURRAY, H. A. *Thematic Apperception Test manual.* Cambridge: Harvard Univer. Press, 1953.

REITMAN, W. R., & ATKINSON, J. W. Some methodological problems in the use of thematic apperceptive measures of human motives. In J. W. Atkinson (Ed.), *Motives in fantasy, action and society.* Princeton, N. J.: Van Nostrand, 1958. Pp. 664–683

RICCIUTI, H. N. *The prediction of academic grades with a projective test of achievement motivation: I. Initial validation studies.* Princeton, N. J.: Educational Testing Service, 1954.

RICCIUTI, H. N., & SADACCA, R. *The prediction of academic grades with a projective test of achievement motivation: II. Cross validation at the high school level.* Princeton, N. J.: Educational Testing Service, 1955.

RIESS, B. F., SCHWARTZ, E. K., & COTTINGHAM, ALICE. An experimental critique of assumptions underlying the Negro version of the TAT. *J. abnorm. soc. Psychol.,* 1950, **45,** 700–709.

ROSEN, B. The achievement syndrome: A psychocultural dimension of social stratification. *Amer. sociol. Rev.,* 1956, **21,** 203–211. (Also in J. W. Atkinson (Ed.), *Motives in fantasy, action and society.* Princeton, N. J.: Van Nostrand, 1958. Pp. 495–509.)

SANFORD, F. H. The use of a projective device in attitude surveying. *Publ. opin. Quart.,* 1950, **14,** 697–709.

SANFORD, F. H., & ROSENSTOCK, J. Projective techniques on the doorstep. *J. abnorm. soc. Psychol.,* 1952, **47,** 3–16.

SCHWARTZ, E. K., RIESS, B. F., & COTTINGHAM, ALICE. Further critical evaluation of the Negro version of the TAT. *J. proj. Tech.,* 1951, **15,** 394–400.

SCOTT, W. A. The avoidance of threatening material in imaginative behavior. *J. abnorm. soc. Psychol.,* 1956, **52,** 338–346.

SNEDECOR, G. W. *Statistical methods.* Ames, Iowa: Iowa State College Press, 1946.

TOMPKINS, S. S., & MINER, J. B. *The Tompkins-Horn Picture Arrangement Test.* New York: Springer, 1958.

VEROFF, J., WILCOX, SUE, & ATKINSON, J. W. The achievement motive in high school and college age women. *J. abnorm. soc. Psychol.,* 1953, **48,** 108-119.

VOGEL, MARGARET. An investigation of the affiliation motive in college age women using low cue strength pictures. Unpublished honors' thesis, Univer. Mich., 1954.

WALKER, E. L., & ATKINSON, J. W. The expression of fear-related motivation in thematic apperception as a function of proximity to an atomic explosion. In J. W. Atkinson (Ed.), *Motives in fantasy, action and society.* Princeton, N. J.: Van Nostrand, 1958. Pp. 143–159.

(Received January 12, 1960)

APPENDIX

APPROXIMATE SAMPLING ERROR DIFFERENCES FOR PERCENTAGES FROM 35% TO 65%

N	700	55	300	200	100	75	50	25
700	5.4–7.6	5.9–8.3	6.9–9.7	8.0–11.2	10.7–14.4	12.1–15.9	14.6–18.9	20.3–25.8
500		6.3–8.8	7.2–10.1	8.4–11.8	11.0–14.8	12.3–16.2	14.8–19.2	20.5–26.1
300			8.2–11.5	9.1–12.7	11.5–15.6	12.9–17.0	15.4–19.7	20.8–26.5
200				10.0–14.0	12.2–16.7	13.4–17.9	15.8–20.7	21.2–27.2
100					14.1–19.0	15.2–20.1	17.3–22.4	22.4–28.6
75						16.2–21.2	18.2–23.5	23.1–29.4
50							20.0–25.7	24.5–31.1
25								28.3–35.8

Note.—The values shown are the differences required for significance (two standard errors) in comparisons of percentages derived from two different subgroups of the current survey. Two values—low and high—are given for each cell. The lower estimates are based on simple random samples. The higher values are based on the computation of individual sampling errors carried out on the current study data, and allow for the departure from simple random sampling in the survey design such as stratification and clustering. The sampling error does not measure the total error involved in specific survey estimates since it does not include non-response and reporting errors.

Vol. 74, No. 13 Whole No. 500, 1960

Psychological Monographs: General and Applied

WATER REGULATION IN THE RAT:

III. THE ARTIFICIAL CONTROL OF THIRST WITH STOMACH LOADS OF WATER AND SODIUM CHLORIDE[1]

LAWRENCE I. O'KELLY AND ROBERT C. BECK[2]
University of Illinois

Although the study of motivational variables in relation to other behavioral phenomena has become a matter of increasing interest, the actual number of ways in which motivation can be controlled is indeed small. In animal experiments under hunger or thirst motivation, procedures have been limited almost exclusively to variations in the techniques of deprivation. A brief description of some of the variants of the "natural" procedure will serve to clarify the need for additional methods of controlling motivational variables. The purpose of this monograph is (*a*) to describe a procedure for "artificially" manipulating the thirst drive of rats by the placement of stomach loads of varying volume and composition, and (*b*) to present a series of standardizing, parametric studies of some of the principle variables involved.

THIRST CONTROL

Natural Methods

Subjects maintained on deprivation schedules differing in duration. The shortcomings of this procedure are several. Under differing hours of water deprivation the *S*s have different degrees of associated hunger, it having been shown that food intake is less when water is absent from the animal's living quarters. Thus, thirst *and* hunger effects are confounded. Further, if, having maintained *S*s on a particular deprivation schedule, one wishes to shift them abruptly to a different schedule, the *S*s may turn out to be somewhat less flexible than the experimenter. The animal adapts to particular maintenance schedules only gradually, and changes in behavior following schedule disruption cannot be attributed solely to changes in "drive strength," for example. Differences in duration of deprivation often must be quite large in order to produce behavioral differences of measurable amount. This means that *S*s may have to be run at irregular times which may fall in different parts of the same or different diurnal cycles, a factor which is usually reflected in greatly increased performance variability. Another serious hazard of deprivational methods arises when they are used for producing drives of high intensity. Long deprivation periods (longer than 36 hours of water lack) inevitably result in debilitation and inanition when continued long enough to be of any experimental value, thus introducing another confounding factor. For example, in situations where performance declines from maxima achieved at intermediate deprivational levels, it has always been difficult to determine whether drive increases continuously with hours of deprivation.

Predrinking. Sometimes all *S*s are maintained on a common schedule, but are allowed to drink different amounts of water prior to testing or training. This, however, has the disadvantage that differential predrinking itself is a variable affecting performance (Bruce, 1937, 1938; Miller, Sampliner, & Woodrow, 1957). Moreover, if an investigator wishes one of several groups to be maintained on a long or severe deprivation schedule, then all *S*s must be maintained on that schedule, with the consequent health risks as noted above.

---

[1] This series of experiments was partially supported by grants from the National Science Foundation and from the Graduate College of the University of Illinois.

Appreciation is expressed to Emanual Lask, Burton Slotnick, and Yasuko Azuma, who assisted in various stages of collecting and analyzing data.

[2] United States Public Health Service Postdoctoral Fellow. Now at Wake Forest College, Winston-Salem, North Carolina.

Running at varying times after drinking. A variation of the above procedure is to maintain all *S*s on the same schedule, but run them at varying intervals after drinking. This has some advantage over a predrinking method, but still does not solve the differential hunger problem since *S*s typically eat just after they have drunk. One can readily shift hours of deprivation simply by running the *S*s at different times after drinking; but if the *S*s have learned to perform at a certain time it is difficult to know whether to interpret the change in performance as being due to differences in drive per se or to the changes in the running schedule.

Maintenance at a percentage of body weight. This deprivation is commonly used with both rats and pigeons for the control of hunger, but does not appear to be well-suited to control of thirst. Whereas a 10-20% weight reduction with food deprivation makes a good active animal for much research, such a water loss could not be tolerated by the rat. Most important, perhaps, is the fact that the effects of a weight-controlling procedure would inevitably be confounded with the effects of the accompanying reduced food intake; so that the loss could not appropriately be attributed to water deficit alone, and the water loss itself could not be differentially estimated.

Artificially Controlled Thirst

Because of the many disadvantages and inconveniences attached to control of thirst solely through duration of water deprivation, we have expended considerable effort on the study of alternative ways of accomplishing the same ends. There are a number of very clear advantages of the artificial methods. All *S*s can be maintained under comparable living schedules, irrespective of any differences in the motivational states they are under during the actual experimental observations. A variety of intensities of thirst can even be induced in the *same* animal on different occasions without change of the daily maintenance schedule. An earlier study (O'Kelly, 1954) showed that the effects of water loads are independent of each other if there is a two-day recovery period between loads; in the experiments to be reported in this paper, the same is true for a variety of NaCl loads, although the length of the recovery period is slightly longer.

Another important advantage of artificial control is that variations in thirst can be obtained and sustained independently of variations in consummatory activity, thus permitting the study of relationships between various parameters of tissue need and behavior without the complications introduced by behavioral consequences of the consummatory activity itself. Whether physiological imbalances vary concomitantly with *drive* is still an important unanswered question (O'Kelly & Heyer, 1951; Towbin, 1949). Wayner and Reimanis (1958) found that running speed increased as a function of time after hypertonic salt injections, but that the amount of water consumed remained constant. Experiments of this type illustrate the advantages of being able to influence separately those variables of tissue disequilibrium from the temporal and spatial supports with which they are inevitably associated in natural deprivation.

Methods of Control

It is well known that drinking can be stimulated artificially by introducing into the animal's blood stream a number of solutions of concentration hypertonic to that of the plasma (i.e., having an effective osmotic pressure greater than that of the plasma). Sodium chloride has been the substance of choice, since, as the normal major extracellular electrolyte, it produces the best combination of drinking enhancement with minimal toxic or other side effects. The solution may be delivered to the plasma by a variety of routes: intravenously, intraperitoneally, or subcutaneously by injection; directly into the hypothalamus or third ventricle by stereotactically placed hypodermic needles; or, finally, introduced into the digestive tract by fistula or esophageal tubing. Since the present monograph is concerned only with use of the stomach loading approach, we will limit our comments on the other techniques to those results that have stemmed from our own previous trials of the subcutaneous injection of saline.

Subcutaneous injection. Proceeding from the findings of Gilman (1937) and of Holmes and Gregersen (1950) that intravenous injections of hypertonic sodium chloride solution increased water intake, Heyer (1951) showed that similar enhancement of drinking in the rat could be induced by subcutaneous injections of 15% NaCl solutions. Further, by varying the amount of NaCl injected, water intakes equivalent to those observed after 12, 24, or 36 hours of water deprivation could be secured. Further standardization of the injection technique was done by Wayner (1953) and Wayner and Reimanis (1958). While drinking can be potentiated by this means, certain disadvantages are characteristic of the procedure. The most serious is the skin damage and pain to animals that may accompany too long a series of repeated injections. This limits the number of days over which unconfounded observations may be made. Also, by the nature of the treatment, it is practicable only to *enhance* drinking from any base reference level, whereas an ideal technique of drive manipulation should permit controlled variation from satiation on up (and possibly down, in the direction of water excess). While drinking potentiation is a function of the total amount of salt, rather than its concentration, drinking inhibition is a function of volume; subcutaneous injections of volumes of sufficient size to limit materially the subsequent water intake would be impracticable to administer.

Stomach loading. For use with rats or other animals of a similar size, it would appear that loading of fluids directly into the stomach via esophageal tubing is the better method. The rat's stomach will accommodate relatively large volumes, thus giving the experimenter a greater latitude of choice in both volume and concentration of administered fluid. Since the rat has no regurgitatory reflex, fluid is not lost during or after placement, as is often the case with hypodermic injection. Also, quite importantly, the route through stomach and intestine is the usual path taken by ingested fluids, and the various steps along the way from stomach to ultimate arrival in the extracellular fluids will more nearly involve normal physiological processes and normal latencies of action.

Administration of fluids in this manner appears to be relatively painless, and it is a matter of usual observation that animals do not develop avoidance habits in relation to the experimenter or the general laboratory situation. As the technique is used in our laboratory, the rats are lightly etherized before loading. This reduces struggling, and the procedure can easily be accomplished with a minimum of difficulty by one person. The time for etherization and loading normally range from 1.5 to 2 minutes. General recovery from the anesthetic is rapid, the animals appearing grossly normal within 5 minutes. There is a small but reliable depression of drinking as a result of exposure to ether (O'Kelly & Weiss, 1955), but this does not appear to interact in any specific manner with load effects.

Previous studies using the stomach loading technique have demonstrated that (*a*) water consumption by 23.5-hour water-deprived rats can be depressed in quantitative fashion by varying volumes of water loads, or can be enhanced by hypertonic NaCl loads (O'Kelly, 1954); (*b*) bar pressing may be similarly enhanced or depressed by such loads (O'Kelly & Falk, 1958); and (*c*) runway performance is affected in a like manner (Solarz, 1958). Thus, the technique is capable of producing behavioral effects consistent with the manipulated need state. Finally, water regulation has been studied in some detail by means of this and related procedures so that such physiological variables as rate of clearance of fluids from the stomach and their intestinal absorption following loading are partially understood (O'Kelly & Falk, 1958).

The studies presented in this monograph are a continuation of the program cited above. They constitute a further extension of our efforts to quantify the control of thirst in the rat by means of stomach loading technique, and to study relations between the need state of tissue dehydration and resulting behavior.

OUTLINE OF GENERAL PROCEDURE

With the exception of various details of management of experimental variables, the procedure used in the drinking studies (Experiments 1-4) was the same for all, and can be outlined first.

Subjects. The *S*s in Experiments 1–4 were experimentally naive male albino rats of the Holtzman strain, from 100 to 130 days of age at the start of the experiments. They had Purina laboratory chow constantly present in their home cages during all phases of the experiments, but once an experiment was started they did not receive water in the home cages until it had been completed.

Adaptation and recovery periods. Prior to any experimental treatment the *S*s were put on a 23.5-hr. water deprivation schedule for 10 days. Each day they were brought from the home cages to the laboratory where they were weighed and then given access to water for a period of 30 min. in individual drinking boxes. On load days they were weighed at the regular time, but drank in special cages, described below. On all interload days, and for a number of days after the final load day, they were weighed and watered just as during adaptation. Thus a check was maintained on the effects of the schedule and treatments, and the general state of their health was closely followed.

Apparatus. The apparatus consisted of drinking boxes, metabolism cages, and water tubes. The drinking boxes, 4 in. by 11 in. by 6 in. high were constructed of wood and had hinged wire tops. The metabolism cages, circular in shape and made of hardware cloth, were 6 in. high, 8 in. in diameter, and rested on double thicknesses of fine copper wire screen set in the mouths of glass funnels. Underneath each funnel was a 50-cc. centrifuge tube for the collection of urine, the feces being screened out. Funnels and screens were cleaned before each session to insure uncontaminated urine. The drinking tubes were 100-cc. gas-measuring tubes, graduated to 0.2 cc., into which were placed drawn glass drinking spouts. The ends of the drinking tubes were inserted into the drinking boxes and metabolism cages about 1 in. above floor level.

Solutions and drinking water. All saline solutions were made with distilled water, using a weight/volume ratio (Pfaffman, Young, Detheir, Richter, & Stellar, 1954). All loads were given at room temperature. All drinking water was local tap water, given at room temperature. Chloride concentration of the load solutions was checked by chemical analysis.

Load day procedure. On load days the animals were weighed at the regular drinking time and the exact volume of load was determined for each animal. Each *S* was then anesthetized lightly by placing it in a museum jar saturated with ether fumes and given its load by stomach tube (urinary catheter), as has been previously described (O'Kelly, 1954). It was then put into its metabolism cage while it recovered from the ether. A 15-min. delay period was used in all the drinking experiments except 4, which was specifically a study of the effects of delay of access to water after loading. At the end of the delay period the water tube was inserted into the cage and water intake and urine output were recorded at 5-min. intervals for 30 min., 15-min. intervals for the next 90 min., and at 30-min. intervals for the final 120 min., a total of 4 hr. At the end of 2 hr. the centrifuge tubes collecting the urine were changed so that separate urinalyses could be made on the first and last halves of the drinking period.

Urinalyses. Immediately following their collection, all urine samples were put into screw-cap vials and refrigerated until they could be analyzed. In the first two experiments analyses were made for chloride content, using the method of Schales and Schales (1941). In the later drinking experiments sodium and potassium content were determined as well, using a Beckman flame photometer.

EXPERIMENTS 1 AND 2: THE EFFECTS OF WATER AND SODIUM CHLORIDE LOADS ON THIRSTY AND SATIATED RATS[3]

Method

Subjects and experimental design, Experiment 1. Sixty rats were used, averaging 361 grams in weight on the first load day. The experiment was designed specifically to study the effects of five different volumes of hypertonic loads and water on the 4-hr. intake of 23.5-hr. deprived rats.

---

[3] R. C. Beck and L. I. O'Kelly. Since the responsibility for design and execution of the various parts of this series varies from experiment to experiment, authorship will be indicated in this fashion.

TABLE 1

VOLUMES, CONCENTRATIONS, AND ABSOLUTE NACL CONTENT OF LOAD SOLUTIONS

Volume % of Body Weight	NaCl Concentration (grams/litre)						
1	0[a]	3.0	*5.8*	*9.1*	*12.0*	*15.3*	*18.5*
2	0		*3.0*	*4.5*	*5.8*	*7.4*	*9.1*
3	0		*2.0*	*3.0*	*3.9*	*4.9*	*5.8*
4	0		1.5	2.3	3.0	3.6	4.5
5	0		1.2	1.7	2.4	3.0	3.6
	Isosals (grams NaCl/kilogram body weight)[b]						
		0.3	0.6	0.9	1.2	1.5	1.8

Note.—All loads were used in Experiment 1. Italicized loads were used in Experiment 2.

[a] "0" indicates tap water load.

[b] Due to slight errors in mixing solutions the isosals are medians for the five different volumes of load.

There were five different NaCl concentrations at each volume. The treatments are summarized in Table 1.

Three characteristics of Table 1 should be pointed out. First, the concentrations at each volume were selected so that the highest concentration would approach a near-lethal dose and so that there would be equal steps between the concentrations at a given volume. The differences depart slightly from equality due to slight variability in mixing the solutions. A lethal dose was estimated as being an absolute amount of salt equal to about two grams per kilogram of body weight. This estimate was based on extrapolation from results in preliminary studies which had included a few actual lethal doses.

Secondly, each of the five concentrations at a given volume represent the same absolute amount of salt as are found in the concentrations at any other load volume. Consequently, we have two ways of studying the effects of concentration of solution on drinking: (*a*) by varying the volume of load and holding the absolute amount of salt constant (reading down each of the columns in Table 1); or, (*b*) by holding the volume constant and varying the amount of salt (reading across each of the rows in Table 1). A given amount of NaCl, in varying volumes of solution, is referred to throughout this paper as an *isosal.* There are in this experiment five *isosals,* the tap water loads not being considered as such.

Finally, the 3.0% concentration was included at the 1% volume so that this concentration would appear with all volumes. It alone does not have an isosal counterpart at all volumes.

The 60 *S*s were divided into two groups of 30, making two simultaneously run replications of the experiment, one in the morning and the other in the afternoon. Each *S* received a different load on each of three separate days, with a four-day recovery period between loads. Treatments were randomly assigned to animals with the restrictions that (*a*) no *S* received the same load twice, and (*b*) each treatment occurred equally often on first, second, and third load days. There was a slight departure from the last restriction in that there were five, rather than six, observations for six of the 31 treatments. Because it was impossible to load and observe more than 10 animals in a half-day, the experiment was staggered, each treatment "day" being divided into three separate daily sessions, with 10 animals loaded in each session.

Subjects and experimental design, Experiment 2. Whereas the first experiment used *S*s 23.5-hr. water-deprived at the time of loading, this study dealt with the effects of such loads on "satiated" rats, i.e., rats in a state of water balance at the time of loading. Since it was deemed desirable to maintain *S*s on a drinking schedule, satiation was achieved on load days by giving them their regular 30-min. drinking period. They were then removed to a common retaining cage and following a 1-hr. delay, a time previously shown to be sufficient for water to clear the stomach and be absorbed from the intestine (O'Kelly, Falk, & Flint, 1958), were loaded and given their 4-hr. drinking period.

Thirty rats were used, weighing an average of 369 grams on the first load day, about 8 grams heavier than the *S*s in Experiment 1. The only difference in design from Experiment 1 was that the hypertonic loads were given at only three volumes (1, 2, and 3% body weight) and no water loads were given. The volumes and concentrations used are those in italics in Table 1. Since each animal was loaded three times, there was a total of 90 loads, 6 *S*s for each of 15 treatments.

Results of Experiments 1 and 2

In order to facilitate comparisons, the results of both experiments are presented simultaneously. Before doing so, however, it is necessary to comment on the statistics. Unless otherwise stated the analyses were three-way analyses of variance: load concentration × load volume × day of loading (first, second, or third). Since the number of *S*s having any particular sequence of treatments was too small to be analyzed, it was assumed that each *S*'s three different treatments were independent of each other, and that the number of independent observations was equal to the number of loads given. In Experiment 1 only the treatments including the five isosals common to all volumes were used in the analyses; water loads were excluded. There being no differences between replications, they were combined. All the treatments in Experiment 2 were included in the analysis. Only significant *F* ratios are reported, these being given at appropriate places in the text.

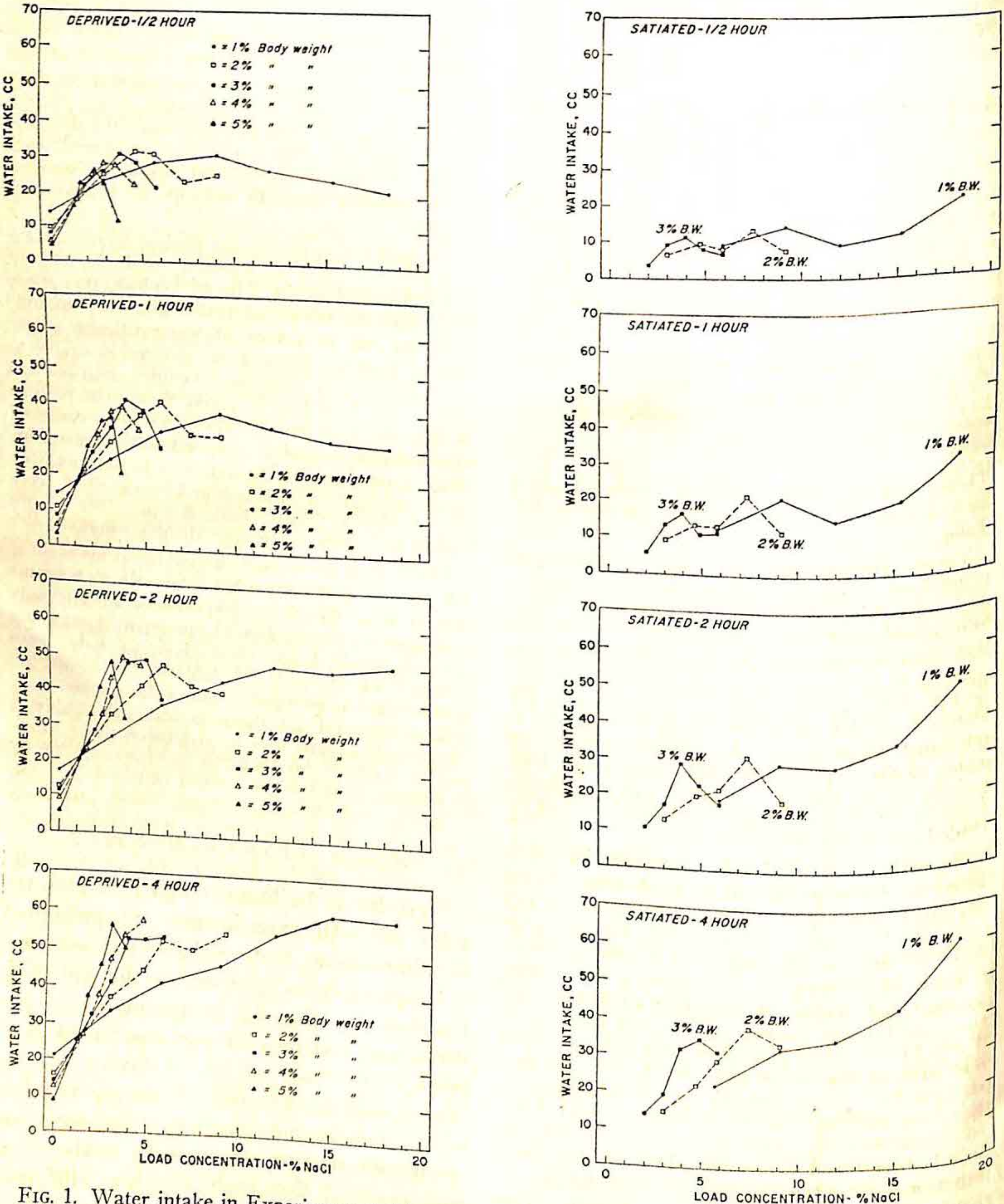


FIG. 1. Water intake in Experiments 1 (deprived) and 2 (satiated) as a function of load volume and concentration at four intervals after access to water; ½, 1, 2, and 4 hours. (Intakes of 23.5-hr. deprived control animals—no experimental treatment—are 19.2, 19.7, 21.2, and 24.7 cc. for these same intervals.)

Water Intake

Figure 1 summarizes the 30-, 60-, 120-, and 240-minute water intakes as a function of volume and concentration of load. For the deprived animals, in the early part of the drinking period (30-60 minutes) drinking increases in a near-linear fashion with load concentration up to some optimal point of concentration, and then declines. With increasing volumes the optima occur with progressively smaller concentrations and in fact occur at each volume with approximately the same amount of salt, between 0.9 and 1.2 grams per kilogram of body weight. The exact optimal salt load is indeterminate from these data, but data from another experiment indicate that the optimal concentration for 2% volume is about 5%, or 1 gram/kilogram of body weight.

The form of the 30-minute intake curves differs between the deprived and satiated *Ss*. The 3% volume curve is similar in the deprived and the satiated groups, but the 2% deviates at the 5.8% concentration, which is lower than the 4.5% when it is expected to be higher. The 1% volume curve shows little patterning at 30 minutes. Individual drinking curves show that the reason for this discrepancy between the satiated and deprived groups is the ability of some of the satiated *S*s to go for long periods without ingesting water. For example, one *S* given a 3% volume of 6% NaCl (the maximum volume and concentration in Experiment 2) drank nothing in four hours. A number of other animals failed to drink in two or three hours.

Total intake for four hours tends to be proportional to the amount of salt loaded in both experiments, except that the most concentrated loads at any volume produced an inversion in this amount of time. With drinking periods longer than four hours, presumably all would be monotonic functions.

Separate analyses of variance were calculated in each experiment for intake during the first two hours, second two hours, and for the total four hours. The results were essentially the same for both experiments, except that there was a significant day effect in Experiment 1, the total mean intakes being 43.7, 46.5, and 50.1 cc. on the successive load days. (This effect is studied more exhaustively in Experiment 3.) The most potent influence was the amount of salt in the load, with volume playing a much smaller role, and with almost no interaction between the three variables. With increasing load volumes there was a corresponding decline in drinking, presumably due to the extra water included in the larger loads, a circumstance requiring the animals to take in a lesser fraction of their water by drinking.

Further analyses were computed for the two experiments jointly, using the 1, 2, and 3% volumes from Experiment 1. The deprivation effect was highly significant during the first two hours ($F = 130.96$, but not during the second two hours ($F < 1.00$). When the amount predrunk by the satiated animals was added to their first two-hour intake, however, the F value for deprivation was reduced literally to zero. The average four-hour intake for the deprived animals at the 1, 2, and 3% volumes was 48.80 cc.; that of the satiated animals, with predrink added, was 47.56 cc. It seems clear that the deprived animals drink first of all to offset their 23.5-hour deficit, then to counteract the salt loads.

The initial high intake rate of the deprived animals as compared to the satiated is demonstrated clearly in Figure 2, which shows the rate of intake over four hours as a function of isosals, with all load volumes combined. These curves, though somewhat stylized due to combining, still reflect accurately the individual volume curves. For the deprived animals, at all volumes, the initial depression of drinking due to high concentration loads is gradually made up throughout the four hours, whereas at the lower concentrations the animals take in their water quickly (30-60 minutes) and drink very little thereafter. The satiated animals drink at a slower rate in the first hour than do the deprived, but with the higher isosal loads they, too, drink steadily throughout the four hours.

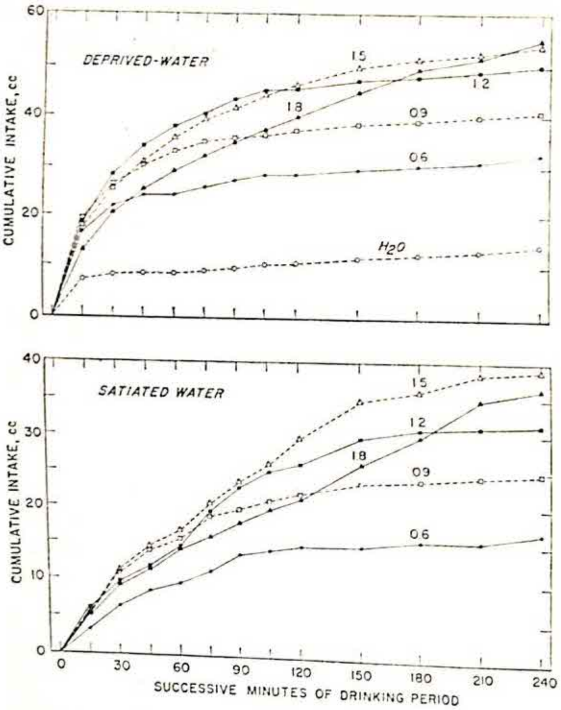


FIG. 2. Four-hour drinking rates as a function of isosal in Experiments 1 (deprived) and 2 (satiated). (Time zero on the abscissa is when the animals were given access to water after loading. For the H_2O loads and each isosal, the rates for all load volumes have been combined.)

Figure 3, from data in Experiment 1, shows the rate curves for five volumes of H_2O load and the five volumes of 3% NaCl load. Water intake following H_2O loading decreases with the magnitude of the load over the entire four hours, but the rates for varying volumes of 3% NaCl show the same general trend as the data in Figure 2, with an increase in intake up to an optimal point during the early part of the drinking period, then a depression with higher volume. At the end of four hours intake is essentially a linear function of load volume. These results give further indication that the amount of salt in the load is the major determiner of intake.

Urine Output

Figure 4 shows the rate of urine output for the deprived and satiated *S*s. As with water intake, the most potent determiner of urine volume is the amount of salt in the load, *F* values of 64.23 and 24.74 being obtained in the two experiments for four-hour urine output. The *F* values for the variable of load volume, on the other hand, were 7.70 (less than .01) in Experiment 1 and 2.32 (*ns*) for Experiment 2. The day effect was not significant in either experiment.

In Experiment 1, urine volume increased with load volume up to 4% body weight during the first two hours, but declined at 5%. For the total four hours, on the other hand, the increase was monotonic with load volume. In Experiment 2, both two-hour and four-hour urine volumes increased from 1% to 2% load volumes, then declined at 3% load volumes.

Generally, however, the most striking effect with respect to urine volume is the almost complete lack of difference between the two experiments. A combined analysis of variance showed that while the satiated animals did excrete a statistically greater amount ($p = .05$) than the deprived during the second two hours, there was no difference for the first two hours or total four

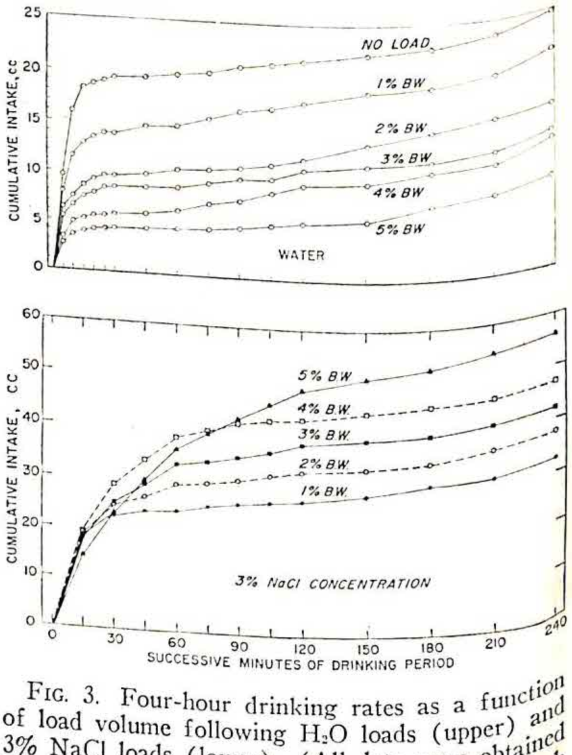


FIG. 3. Four-hour drinking rates as a function of load volume following H_2O loads (upper) and 3% NaCl loads (lower). (All data were obtained from Experiment 1, the number of *S*s for each curve being six.)

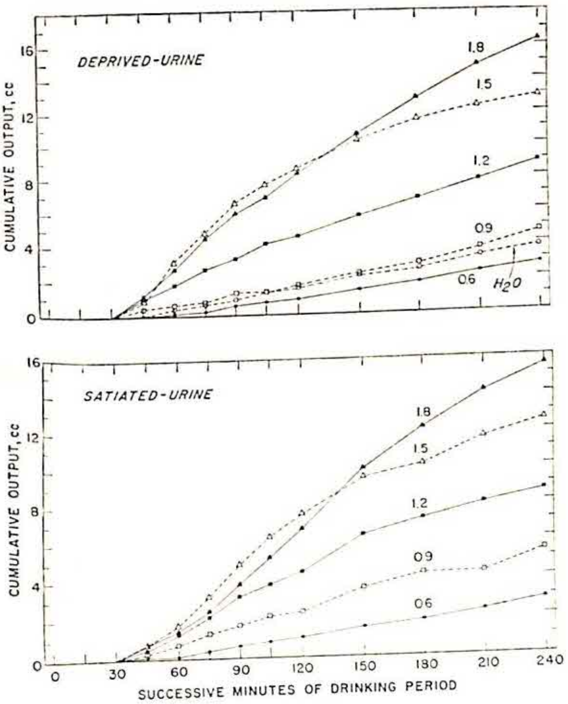


FIG. 4. Four-hour urine rates for Experiments 1 and 2 as a function of isosal. (The initial higher rate with the 1.5 gm/kgm isosal than with the 1.8 appears to be due to the animals with the 1.5 load having a diarrheic load effect earlier than those with the 1.8 load. In Experiment 1 the antidiuretic effect of low concentration loads is seen where the 0.6 isosal load is followed by lesser urination than the H_2O load.)

hours. The predrink and hour's delay between predrink and loading of the satiated animals evidently does not influence delay of onset of urination after loading, rate of output, or total urine flow.

We find then, with respect to both water intake and urine volume that there are no differences in the deprived and satiated groups other than the extra water needed by the deprived animals to offset their previously incurred 23.5-hour deficit.

The upper part of Figure 5 shows the urinary chloride concentrations for both experiments as a function of isosal. Since urines were collected separately for the first two hours and second two hours, this "half" variance was included in the analysis. Once again, the isosal effect is very large, and in Experiment 2 it is the only significant variance. An unexpected occurrence was the highly significant difference between halves in Experiment 1, where the second-half urine is much more dilute than that from the first half. This effect did not occur in Experiment 2, however, and a combined analysis of variance for the two experiments shows an *F* value of 24.86 between deprived and satiated in the second half.

The animals have two urinary means of ridding themselves of the imposed NaCl load. They can excrete a larger volume of more dilute urine or a smaller volume of more concentrated urine. As Figure 5 shows, the animals have a fairly constant increase in total urine volume, as a function of isosal, over the four-hour period. Hence this mechanism of chloride excretion (volume) is being used throughout the entire range of isosals. The concentration of urine, on the other hand, increases only up to the second or third isosal, then remains at a level of about 2% (although individual animals were able to concentrate their urine as high as 3%). The lower part of Figure 5 shows the product of concentration and

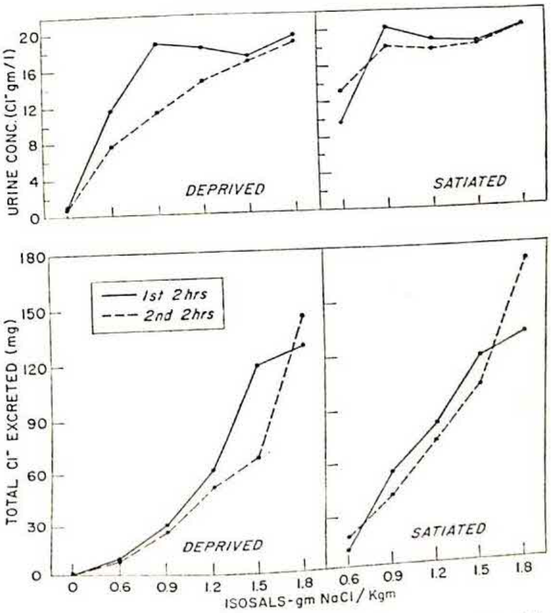


FIG. 5. Urine concentration in grams of chloride per liter (upper), and total milligrams of chloride excreted (lower) as a function of isosal in Experiments 1 and 2.

volume, the total amount of chloride excreted, in milligrams. Whereas the rats may have reached an upper concentration threshold in both halves of the drinking period, they apparently had not reached a volume limit, nor a limit for chloride excretion in the second half. Otherwise, there would have been some asymptotic limit approached for chloride output in the second two hours, such as is indicated in the first two hours. Apparently, there could have been a still greater volume of urine produced had there been more concentrated loads given. Load concentrations above those in this experiment run dangerously close to lethal dosages, however, so it may be that urine volume would increase right up to the final lethal dose.

In relation to the usefulness of various loads for behavioral experiments, it may be noted that the occurrence of diarrhea following loading is of some practical interest. Diarrhea occurred commonly at the two highest isosals in Experiment 1 (about 40% of the cases), much more frequently than in Experiment 2 probably due to the presence of more water within the intestinal lumen. The number of cases, by increasing isosals, was 0, 3, 6, 13, and 12 in Experiment 1 and 0, 0, 2, 4, and 1 in Experiment 2. A test of the difference in proportions at the two highest isosals was significant beyond the .01 level. Using only the three lowest volume loads in Experiment 1 for this comparison, the significance level was between .05 and .10. In Experiment 1 there were about an equal number of diarrheas at all load volumes (range: 6-9), except at 1% where there were only 2. In Experiment 2, the distribution was 2, 4, and 1, with increasing volume of load.

Interload Day Weight and Drinking Changes

Weight and drinking changes on days after loading are shown in Figure 6. Since there was no regular drinking period on load days in Experiment 1, the day before loading was taken as the baseline. The satiated animals have an extra recovery day since the second and third load days are "recovery" days from the previous loads.

In both experiments the increase in weight following loading reflects the total intake on the previous load day. The weight changes increased with increasing isosals up to the fourth (1.5 gm/kgm), then declined sharply at the fifth. In Experiment 1 there was an overall correlation of .51 between load day intake and weight change on the day after. Water loads, however, had no effect on weight. In each experiment, separate analyses of variance were computed for weight and drinking changes on each post-load day until there were no significant effects. *T* tests were computed for the difference between baseline days and days after, using all observations, irrespective of treatment (H_2O loads excluded). The main difference between the satiated and deprived animals is that the deprived were still significantly overweight at the end of four days, while the satiated animals returned to their preload weights in this period of time. As expected, concentration was the most important variable (a significant effect until the third post-load day), with some small influence of volume. The greater regularity of the curves for deprived *S*s in Figure 6 may simply be a matter of reliability since there were more *S*s in the first experiment, but it more likely corresponds to the greater regularity of intake on load days.

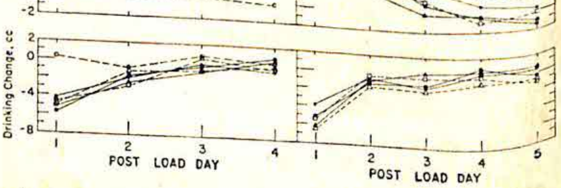


FIG. 6. Weight and drinking changes on days after loading as a function of isosal.

Discussion

The first major result of the present experiment is that drinking, urine volume, and urine concentration are mainly a function of the amount of salt loaded. This agrees with the findings of Wagner and Reimanis (1958) for subcutaneous salt injections. Load volume effects, while significant, are comparatively slight and seem to be due to the fact that, since higher volume loads depart less from isotonicity than small volume loads, they require the animals to drink less water to offset the experimentally-produced deficit.

The second critical result is that salt loads can depress drinking, as well as potentiate it. Thus, whereas the lower concentration loads produced immediate drinking at a high rate (due partly, of course, to the 23.5-hour deficit), the higher concentration loads produce a lower but more persistent rate of drinking. This suggests (with supporting evidence from Experiment 6) that the higher concentration loads are not well-suited for use in behavioral experiments, at least with the 15-minute delay after loading. Direct observation of the stomachs and intestines of animals loaded with high concentration solutions shows them to be very distended, filled with osmotically-induced water from the tissues. Overall activity is considerably reduced. The occurrence of diarrhea serves as a check on this observation.

It would seem a good guess that the highest concentration load practical for the control of behavior would be one at, or slightly below, the optimal load for potentiating drinking in a 30-minute test period (see Figure 1). The loads up to this point tend to show an increment of drinking which is approximately linear and drinking drops off rapidly after about an hour (Figure 2). In other words, these loads seem to produce an immediate change in water deficit which is offset within a minimal time.

The great similarity between the deprived and satiated animals is important with respect to the physiology of salt-load regulation and to the management of behavioral experiments. The deprived *S*s seem to show more regularity than the satiated, however, as shown in their more orderly drinking curves. While larger numbers of animals for the treatments under the two conditions might produce completely analogous drinking results, the lesser variability of the deprived group with only a few animals per condition is just a point in its favor. One of the sources of greater variability for the satiated animals is the fact that they can manage the load either by drinking or by excreting. Some of the satiated animals seemed able to utilize body water very effectively for kidney function during the first part of the drinking period and took in their "environmental water" later in the period. The deprived animals, of course, do not have as much tissue water to expend in kidney regulation at the time they come into the test period.

EXPERIMENT 3: THE EFFECTS OF REPEATED WATER AND NACL LOADS ON THIRSTY RATS[4]

This experiment was designed to study the cumulative effects of six repeated load treatments, using a six-day intertreatment interval rather than the four-day interval of Experiments 1 and 2. For reasons just discussed, concentration values which were on the accelerated portion of the 30-minute intake curve (Figure 1) were selected. Of the five volumes in Experiment 1, the 2% body weight volume seemed to give the greatest amount of drinking. The concentrations used in this volume did not appear to be such as to overly irritate the digestive tract, and the loads were of small enough volume to be readily tubed into the animals. The decision to study thirsty animals rather than satiated was guided by two considerations: the greater regularity of results with the deprived animals, and the possibility of either enhancing or depressing the drinking from a 23.5-hour deprivation base line.

---

[4] R. C. Beck and L. I. O'Kelly.

Method

Since the optimal concentration value for a 2% load appeared to be about 6% (in terms of maximal potentiation of drinking), it was decided to use concentrations of 0 (tap water), 2.0, 4.0, and 6.0% NaCl. Initially there were 10 animals in each group, but in the course of the experiment five were lost due to death or treatment error, so that 35 finally completed the experiment. There were nine subjects in three of the groups and eight in the fourth; the weights of all *S*s averaged 363 grams on the first load day. All were male Holtzman albino rats about 100 days old. Following adaptation as previously described, each *S* was given the same load once a week over a period of six weeks. Methods of collecting drinking and urine data were the same as in the previous experiments.

Results and Discussion

Water Intake

Figure 7 shows the rate of water intake for the four different load concentrations on the first load day. We see here that our choice of 6% NaCl as the "optimal" load for potentiating drinking was incorrect, since these animals actually drank less than those loaded with 4% during the early part of the drinking period. As a check on what the "true" optimum might be, a different group of 10 animals was given a 2% volume load of 5% NaCl. These *S*s, in 30 minutes, drank a mean of 37.2 cc., which was above that previously obtained with any load. After four hours of drinking, however, intake (57.9 cc.) was so far above the extrapolation from lower concentrations as to suggest that concentration was not the only variable operating (see Figure 21). It is clear, however, that the optimal load concentration is closer to 5% than to 6% NaCl.

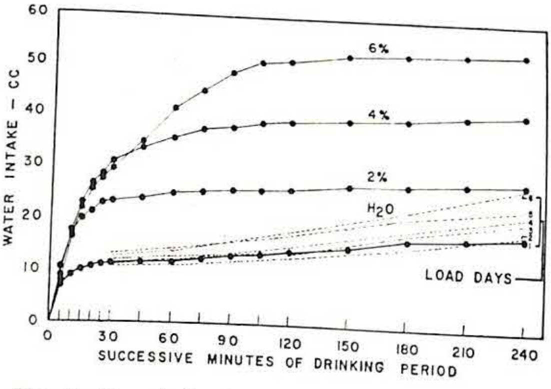


FIG. 7. Cumulative intake for all groups on Load Day 1 (circles) and cumulative intake from 30 minutes on for H_2O loads on Load Days 2 through 6 (dotted lines). (By the end of 4 hr. all H_2O load day intakes are ordered by day: 1-6. The results were similar for the other loads, differing only in absolute values.)

The broken lines in Figure 7 show the rate of intake following H_2O loading on the second through sixth load days. There are no large differences in the first hour or so, but by two hours there is a clear separation of rates, and by two and a half hours intake varies as a function of load day, only the second day being out of order. The results with this particular load concentration are representative of those obtained with the other three concentrations. As Figure 8 shows, total four-hour intake increased equally for all groups over load days.

There was virtually no overlap of individual *S*'s intakes between groups on any given load day. Thus, almost all water-loaded animals drank less than 2% animals, who in turn ingested less than the 4% group, and so on. The differential treatments thus were highly reliable, and there were no observable ill effects to the animals' health occasioned by the repetition of loads.

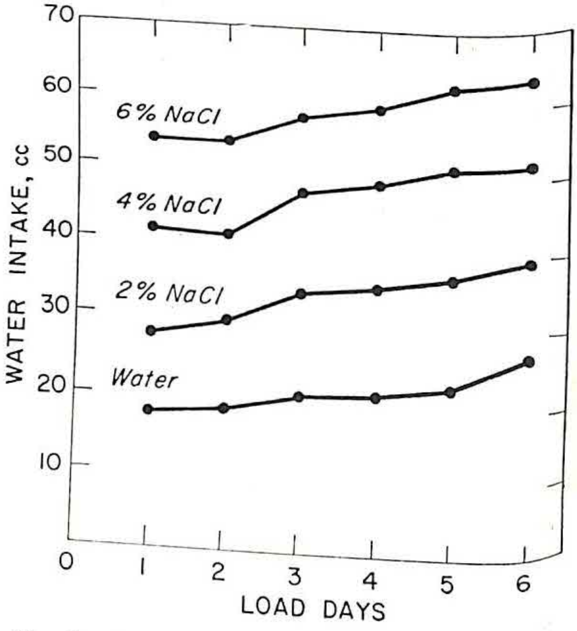


FIG. 8. Four-hour water intake for all groups on successive load days.

The within-group reliability was high also, as indicated by rank-order correlations for total intake from one load day to the next. These correlations averaged between .50 and .70 for adjacent treatment days, and then declined toward zero with an increasing number of treatment days separating the days correlated. Since the groups were small, and were taken from a very homogenous population, relatively minor changes in intake would have a large influence on the size of the correlations. For most purposes the good separation between treatments is the most important consideration. The size of the correlations for successive load days suggests that with larger numbers of *S*s, individual differences could be successfully studied.

Urine Volume and Concentration

The day-to-day changes in four-hour urine output are shown in Figure 9. These changes are of the same sort as occurred with water intake, tending to increase independently of a particular treatment. There is, in fact, a good correlation between intake and urine volume on each load day. The median correlations for the six load days were .74, .70, .25 and .64 for the 0, 2, 4, and 6% loads, respectively.

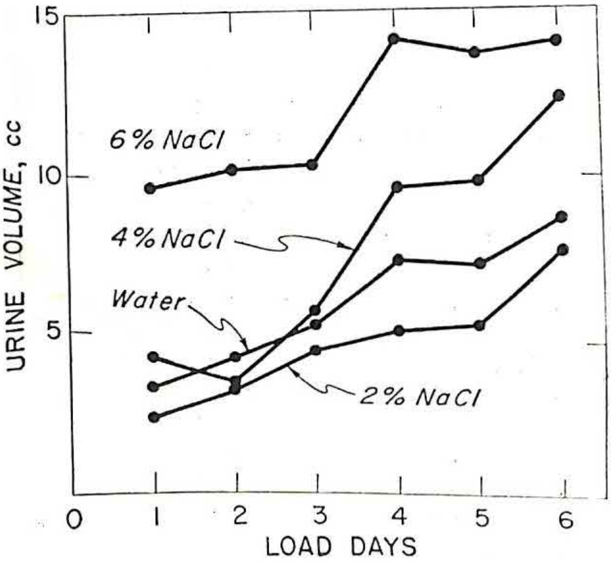


FIG. 9. Four-hour urine volume for all groups on successive load days.

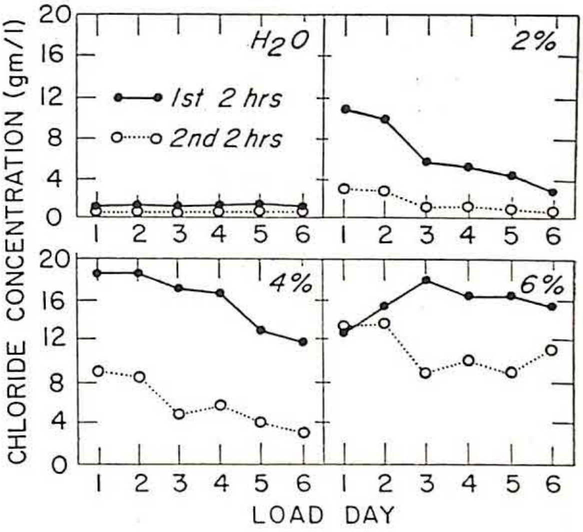


FIG. 10. Chloride concentration during first and second halves of the four-hour drinking period on successive load days.

In Figure 10 the concentrations of urinary chlorides are shown for each of the load days. Whereas chloride concentration tends to decline from day to day, total chloride excreted remains relatively constant or increases slightly. Median chloride excretion in milligrams for the first and last three load days were, for the four groups: 2.3 and 4.0, 14.2 and 13.0, 43.5 and 65.3, and 112.0 and 121.8, respectively. The increase in urine volume offset the decline in concentration of urine, and the net efficiency of chloride excretion remained as great at the end as at the beginning of the experiment. The volume increase in urine parallels the increase in water intake on successive load days, but the direction of causal relationship is not determinable from the present data. It is possible that the repeated dehydration stress of the salt loads, superimposed on a chronic negative water balance maintained by the 23.5-hour deprivation schedule progressively reduces the efficiency of the hormonal regulation of volume excretion, perhaps through exhaustion of antidiuretic hormone. This can only be decided, however, by future experimentation.

Excretion of sodium and potassium through load days was similar to that of chloride. Potassium loss increased as a function of load concentration, presumably an obligatory output stimulated by the sodium of the load (Gamble, 1958, pp. 126–128), but not nearly to the same extent of the sodium or chloride increase. There were no chemical or behavioral indications that potassium excretion was at a dangerously high level under any of the treatment conditions. Reliability correlations for urine volume, ion concentration, and for total ion excreted were consistently positive, but showed greater variability than did those for water intake.

Weight and Drinking on Interload Days

Since weight and drinking changes were shown in some detail in Experiments 1 and 2, it is sufficient to report that the present experiment confirmed the earlier findings. These changes are a function of load concentration and are quite stable from one load day to another.

Daily water intake returns to normal within two or three days after loading, but weight does not necessarily do so. Figure 11 shows the mean weights of the four groups during adaptation, on the six successive load days, and on the final day of the experiment. Although weights did not return to their previous level after loading, there were no permanent differential effects of the different loads. This is in spite of the fact that the immediate changes varied from a 2- or 3-gram decrement on the day following tap water loads to a 20-gram increment following 6% loads. Since the slope of the weight increment curve over the first few load days is the same as the slope over the last few adaptation days, prior to any treatment, it appears that the weight increases are due to normal growth, and are independent of the influence of our load variables.

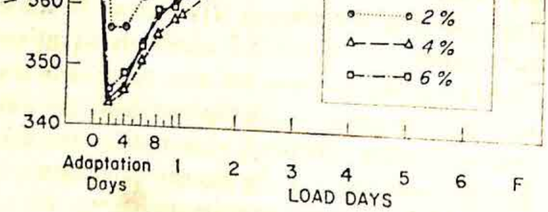


FIG. 11. Mean weights of load groups on alternate adaptation days, each load day, and the final day of the experiment (F). (The interload weight changes—weight increments as a function of load concentration—are not shown.)

The regular 30-minute intake on days prior to each load day was essentially unchanged throughout the course of the experiment, a further indication of good recovery from the loads. Considering all four groups separately on each of the days prior to a loading, and on the final day of the experiment, the mean intake of the groups ranged from 18.4 cc. to 21.9 cc., these means being distributed around a grand mean of 20.4. The mean intake of all animals on the last day of adaptation was 20.06 and on the final day of the experiment it was 20.95 cc. The animals obviously were not drinking appreciably more or less on the deprivation schedule at the end of the experiment than they were at the end of 10 days' adaptation, and intergroup variability was practically nil.

EXPERIMENT 4: DELAY OF ACCESS TO WATER AFTER STOMACH LOADING[5]

One of the important variables in the study of salt loads would appear to be the delay between loading and the *S*s' access to water. While Wayner and Reimanis (1958) found that following hypertonic subcutaneous NaCl injections drinking was independent of delays up to four hours, this may or may not be true for water consumption after stomach loads. Wayner reported no data for urine, but we have shown that both urine volume and urine concentration are functions of load concentration. To the degree that an animal can excrete salt by concentrating its urine before access to water it might need to ingest less water (as was the case with some of the satiated load animals in Experiment 2); and if stomach loading is used in behavioral experiments the consequences of delays between loading and drinking (or training and testing of performance) should be understood.

---

[5] L. I. O'Kelly, R. C. Beck, and Lowell T. Crow.

Method

The *S*s were 120 rats, divided into 24 subgroups of five each and averaging 378 grams in weight. A random groups factorial design was used, varying 2% body weight volumes of 0, 2, 4, and 6% NaCl with delays of 60, 90, 120, 150, 180, and 240 min. between loading and drinking. Each *S* was given the same load and delay twice, with a six-day recovery period after each load. The animals were delayed in the metabolism cages following loading; urine was collected and analyzed from the delay period, as well as the first and second halves of the 4-hr. drinking period.

Results

Water Intake

Figure 12 shows the total four-hour intake as a function of load concentration and delay. These data were taken from the first load day only, but the functions were essentially the same for the second load day (as was true for all of the variables), differing only in the overall magnitude of intake. An analysis of variance showed that load concentration, delay, and load day variables all exerted significant effects on intake far beyond the .001 level. The delay × concentration interaction effect was just significant at the .01 level.

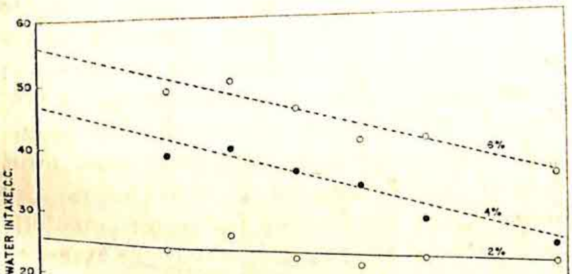


FIG. 12. Four-hour water intake on Load Day 1 as a function of delay of access to water following loading. (The lines are least squares fits to the points.)

The amount of water ingested in four hours showed a somewhat different relation between the load treatments than did either the 30-minute or 60-minute intake. Whereas the decline in consumption as a function of delay shows slightly after a half-hour of drinking and markedly after an hour, even at the 60-minute delay, the 4% and 6% groups drank the same amount in the first 30 minutes (cf. Figure 7). At neither 30 nor 60 minutes after access to water was there any separation between the intakes of the 2% and 4% NaCl groups delayed for four hours.

Weight and Drinking Changes on Interload Days

The changes in weight on the day following loading also declined as a function of delay, as did the 30-minute water intake on the day after loading. This, as in the previous experiments, seems to be a function of total intake on load days. The animals recover their weights and return to regular drinking well within the one-week recovery period, as was found in Experiment 3.

Urine Output

Figure 13 shows urine output during the delay intervals and for the total period of recording (delay plus drinking period). Urine output during the delay period proceeded at a constant rate whose slope was determined by load concentration, but total urine volume was very little affected by the length of the delay interval. The animals excreted about the same volume in five hours as in eight hours. To determine more specifically the effects of access to water on urination, the volume of excretion for the various subgroups was examined at 240 minutes after loading. At this time one group had not yet been given access to water and the other groups had been drinking for various times up to three hours. There were no differences among the six delay groups at each concentration, indicating that urine flow follows a course that

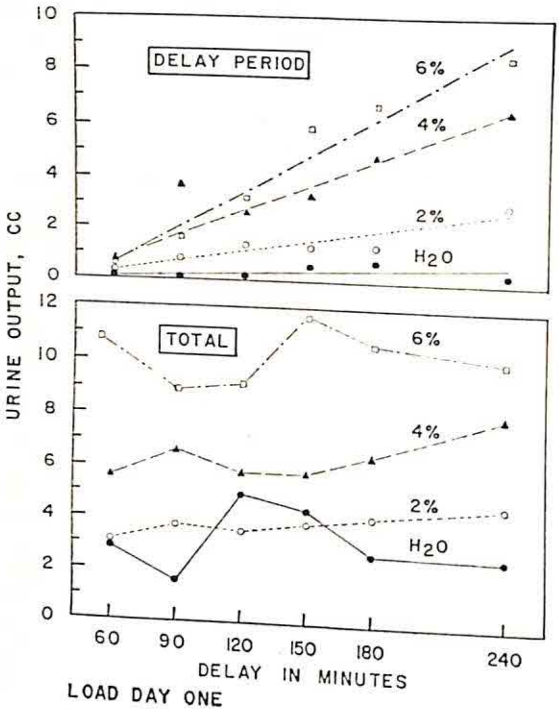


FIG. 13. Mean urine output during delay intervals on Load Day 1 (upper) and total urine volume (delay plus four-hour drinking period, lower).

is little, if any, influenced by whether the animal is drinking or not.

Urine concentration, on the other hand, is strongly affected by the animals' access to water. Chlorides, sodium, and potassium all increase in the urine as a function of delay, as well as of load concentration, as shown in Figure 14, which represents chloride excretion. The absolute amount of sodium and potassium excreted was less than that of chloride, but the functions were of much the same form. Concentration was highest during the delay period (about the same for all the NaCl load concentrations) and progressively declined during the first and second halves of the drinking period.

Discussion

The effects of delay on water intake, etc. are directly attributable to the fact that the *S*s excrete a more concentrated urine and hence a greater proportion of the ionic load if they do not have access to water. Drinking dilutes the fluid load so that it is more readily absorbed. Being diluted, the concentration differential between load and body fluids is lower and does not require as drastic an excretory correction.

These results become more meaningful when compared with those from Experiments 1 and 2. The total intake of the *S*s in those two experiments was the same, and urine volume was the same, but over the range of concentrations used in the present experiment the satiated *S*s put out a more concentrated urine in the last two hours than did the deprived. Since the deprived *S*s immediately drank a great deal more water after loading than did the satiated, the explanation of the urine concentration difference (whatever the exact mechanism may be) would seem to be the same as in the present experiment.

The data obtained in this experiment do not support the results of Wayner and Reimanis (1958), again probably due to differences in excretory control of the imposed loads. Since hypertonic solutions induce movement of body water toward them, in Wayner's experiment the water was initially moving toward the interstitial space under the skin of the back, whereas in our experiment the water went into the stomach and intestine. In this latter case, the absorptive surface is much greater, and corrective

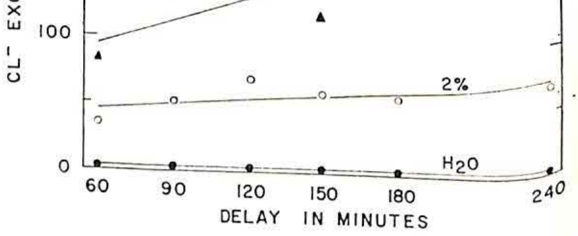


FIG. 14. Total chloride excretion on Load Day 1 as a function of delay. (Whereas urine volume was not necessarily increased by length of delay interval, chloride excretion clearly is for all loads except H_2O.)

filtration by the kidney would be more rapid.

In spite of the various changes that occur as a function of load-drink delay, the effects of concentration on drinking are maintained at all delays, with the exceptions of the 2% and 4% loads at the four-hour delay. With short delays there is no difference between the 4% and 6% loads in the first 30 to 60 minutes of drinking, but this is a temporary inhibition, not a differential management of the loads.

EXPERIMENT 5: A FURTHER STUDY OF GASTROINTESTINAL EXCHANGE RATES OF WATER AND SODIUM CHLORIDE IN THIRSTY RATS[6]

In an earlier experiment (O'Kelly, et al, 1958) the gastrointestinal exchange rates of 3% body weight volumes of water and of 0.5, 0.87, and 3.0% NaCl were determined. The rates were studied up to a period of 35 minutes after loading. These rates were related to bar pressing following loads (O'Kelly & Falk, 1958) and to runway behavior (Solarz, 1958). Other experimenters have indicated the importance of post-ingestive factors in regulating drinking inhibition and other aspects of motivated behavior (Miller & Kesson, 1954; Miller, et al., 1957; Towbin, 1949). If we wish to control motivation by the "artificial" means presented in the previous experiments it is of some importance to know the rates at which the loads move through stomach and intestine and the amounts and distributions of the fluids accumulated in the gut by the osmotic action of the hypertonic loads. Accordingly, the purpose of this experiment was determination of gastrointestinal exchange rates for the 2% body weight volumes of water, 2.0, 4.0, and 6.0% NaCl used in the previous experiments.

Method

Subjects and procedure. The *S*s were 285 male albino rats of the Holtzman strain, varying in age from about 120 to 200 days. All of them had been used in earlier experiments of various kinds, but were readapted to a 23.5-hr. deprivation schedule for at least 10 days before loading and gut determinations were made. The data were gradually collected over the course of a year as the animals became available.

The procedures are described in detail elsewhere (O'Kelly, et al, 1958). Animals are stomach loaded with measured volumes of fluid. After varying delays they are sacrificed, and the stomach, intestine, and cecum are clamped off and separately removed. The fluid content of each component is determined by obtaining its weight before and after oven drying to constant weight, with a correction made for inherent tissue water. From the measured volumes of fluid recovered from stomach, intestine, and cecum after the various delays, the rate of movement of fluid through these compartments may be determined.

Experimental design. The experiment called for 24 groups of 10 *S*s each, given one of the four load concentrations at each of six delay periods: 15, 30, 60, 120, 180, and 240 min. Because the intestinal absorption curve for 6% NaCl had not reached an asymptotic level at 240 min., 5 animals with water loads and 10 animals with each of the NaCl loads were sacrificed after 300-min. delay. Additional 6% NaCl animals were run at 360 and 420 min., six at the former and four at the latter delay.

Results and Discussion

The results are summarized graphically, each point being the mean of all animals sacrificed under the particular delay interval indicated. Figure 15 shows the fluid recovered from the stomach after the various delays. Rate of fluid clearance is an orderly function of concentration. Water and 2% NaCl clear exponentially. After an initial period of stabilization at 70–90% of load

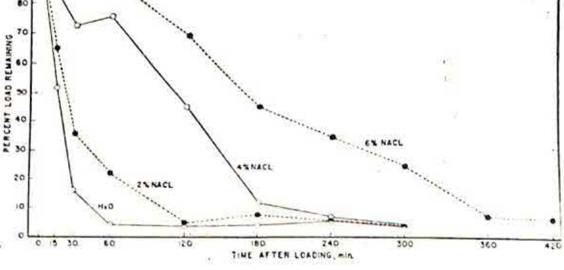


FIG. 15. Percent load volume remaining in stomach at various delays after loading.

---

[6] L. I. O'Kelly and R. C. Beck.

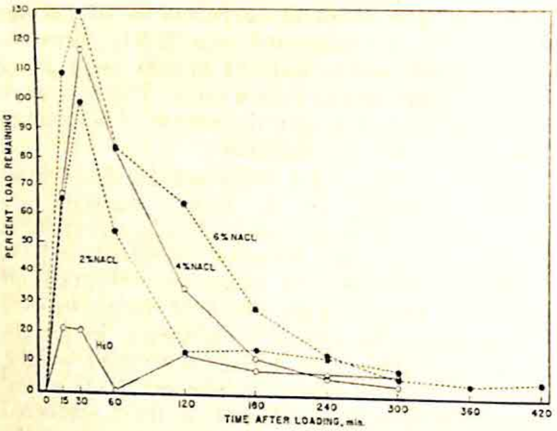


FIG. 16. Percent load volume remaining in the small intestine.

volume, the 4% and 6% NaCl solutions seem to clear in a similar fashion.

Figure 16 illustrates the time course of filling and emptying of the small intestine. Water manifests a small net gain of stomach clearance over absorption in the first 30 minutes after loading, but transport to cecum and absorption have emptied the small intestine of the water-loaded animals by 60 minutes. Accumulation of fluids from the hypertonic loads is, by 30 minutes, directly proportional to the NaCl concentration of the load, as is subsequent absorption and cecal flow. Intestinal distention by the hypertonic loads is maximal at 30 minutes and declines rapidly thereafter.

Accumulation and dissipation of fluid in the cecum is shown in Figure 17. Maximum fluid recovery occurred at 60 minutes post-loading, and cecal emptying is inversely proportional to load NaCl concentration. Since no diarrhea was observed during the delay periods, we assume that the major portion of the fluid remaining in the cecum is subsequently absorbed.

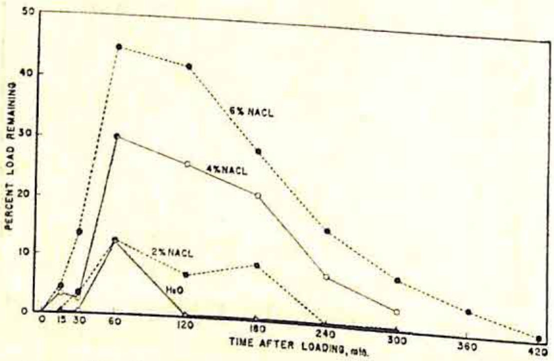


FIG. 17. Percent load volume remaining in the cecum.

Figure 18 shows the percentage of load which is absorbed from the intestine at the several delays. For all the hypertonic loads there is an initial period of negative absorption for some time after loading—i.e., the amount of fluid in the intestine is greater than that given in the load. The rate of negative absorption is a direct function of load concentration, but the inflection point for all loads is at 30 minutes. The 6% load does not return to its initial volume until about three hours after loading, and 88% absorption is not achieved until seven hours after loading.

These curves suggest that in behavioral experiments it would be best to wait at least an hour, and preferably two hours after loading with hypertonic NaCl solutions before requiring the *S*s to perform. Otherwise, the fluid content of the intestine will be rapidly increasing, with whatever effects this might have on performance—probably inhibiting. This problem will be discussed again after we have presented data from Experiment 6.

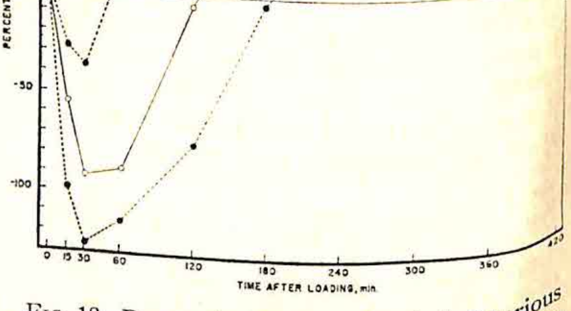


FIG. 18. Percent load volume absorbed at various delay times after loading. (Negative absorption indicates a recovery volume greater than the volume loaded.)

EXPERIMENT 6: ARTIFICIALLY CONTROLLED THIRST AND WATER REINFORCED RUNWAY BEHAVIOR[7]

O'Kelly and Falk (1958) showed that when thirsty rats were loaded with water that water-reinforced bar pressing on a continuous reinforcement schedule was depressed in direct proportion to the load volume, and that it increased with various low concentration NaCl loads. Solarz (1958) found that running slows down as a water load has progressively more time to be absorbed before running is permitted. The present experiment was concerned with the problem: as a function of load concentration and runway length, how many highly massed, water-reinforced trials will a rat run in a single session? This question concerns both the effects of hydration and effort on performance and the methodology of using the loading technique in behavioral experiments.

Method

Subjects. The *S*s were 52 male albino rats of the Holtzman strain, 120-160 days of age at the beginning of the experiment. They had previously been used in other experiments, but had no prior experience with either runways or water deprivation.

Apparatus. A 2-ft. enclosed alleyway, to which a 4-ft. extension could be attached, was used. The delay between raising the start-box door and the *S* breaking a photocell beam 6 in. in front of the door (starting time) and the time between crossing the first beam and a second one 6 in. in front of the goal-box entrance (running time) were automatically recorded. When a water reinforcement of 40 licks was taken from the drinking tube at the back of the goal box, the tube was automatically withdrawn. The amount of water per lick was estimated from preliminary tests to be about .005 cc., or about 0.2 cc. per reinforcement.

Experimental design. Adaptation and runway training: The animals were adapted to a 23.5-hr. deprivation schedule and had two trials daily in the runway for a period of 10 days prior to testing. Each day they were weighed, run, watered for 30 min., then returned to the home cages where they lived in squads of four.

Test day: On Day 11 each member of a squad of four was loaded and after a 25-min. delay was run to its regular water reinforcement until it reached a criterion of three consecutive 60-sec. latencies in leaving the start box. If *S* did not leave the start box within 60 sec. or spent 60 sec. in the alley, he was removed and put into the goal box by *E* to determine if he would drink. Whenever *S* was in the goal box he was removed either just as soon as the tube was pulled out or after 60 sec. if he did not drink. Following removal from the goal box *S* was returned directly to the start box and the next trial was begun in about 15 sec., just enough time to allow *E* to record the times for the previous trial and to reset the clocks. When *S* reached criterion he was removed to the regular drinking box and allowed 60-70 min. of drinking.

For the next four days after testing the *S*s were maintained on the same 23.5-hr. deprivation schedule as during adaptation. On the fifth day they were loaded with the same solution as on the test day, and after a 30-min. delay were sacrificed. Their guts were processed as described in Experiment 5 (they provided the data for the 30-min. delay point in that experiment).

Results

Test Day Performance

Trials to criterion. Table 2 gives the number of trials to criterion for each of the *S*s. These are the number of times each *S* left the start box, whether or not he drank. An analysis of variance showed both the overall concentration variable and runway length to be significant beyond the .01 level of confidence. Individual comparisons of subgroups showed the significance of these functions to be due mainly to the relatively low scores of the H_2O and 6% groups, as opposed to the equally high scores of the 2% and 4% groups.

Water intake. On almost all trials when an animal drank in the goal box the full reinforcement was consumed, making it possible to plot the relationship between the amount of water received at the time the *S* reached criterion (load plus goal-box intake) and total intake (the hour's drink after criterion added). This is shown in Figure 19, which also has the number of reinforcements plotted on the right-hand ordinate. Whereas the function for goal-box intake is parabolic in form, total intake is a monotonic function with no significant differences between the two runway groups.

---

[7] R. C. Beck and L. I. O'Kelly.

TABLE 2

SCORES FOR TOTAL TRIALS RUN TO CRITERION

Runway Length		Load Concentration (% NaCl)			
		0	2	4	6
2 feet		7	40	55	5
		30	42	70	18
		30	55	80	25
		30	67	86	39
		31	79	90	141
		47	81	135	
			94		
	N	6	7	6	5
	M	29.2	65.4	86.0	45.6
	Md	30	67	83	25
	SD	11.7	19.1	24.7	48.9
6 feet		3	3	4	0
		6	10	4	5
		8	25	34	7
		12	42	46	9
		17	55	63	22
		29	78	84	37
		30	84	90	40
	N	7	7	7	7
	M	15.0	42.4	46.4	17.1
	Md	12	52	46	9
	SD	10.1	29.4	32.4	14.9

Note—All cell entries initially had seven *S*s but two were lost to illness, one died during loading, and one was discarded during training for repeated failure to run. The single lowest score for all concentrations with the 6-ft. runway were obtained in the same test session.

The two parabolic functions for number of reinforcements therefore reflect the effects of the different loads and runways on trials to criterion.

The goal-box drinking of the water and 6% groups was somewhat more affected than running; there were many trials when particular *S*s ran to the goal-box entrance but failed to drink. Thus, all the H_2O *S*s combined had a total of 46 runs without drinking and the 6% had 83, as opposed to 8 and 18 for the 2% and 4%. The running of the 6% *S*s was often very peculiar, reminiscent of a fraternity pledge pushing a peanut down the sidewalk with his nose. Such animals, in either runway, often ran to the goal-box entrance, then sat for 60 seconds until put into the goal box, where they would not drink. These same animals, on the other hand, drank immediately when put into the drinking box. The only apparent reason for this failure to drink in the runway was that the tube was slightly recessed, whereas in the drinking box the tube extended about a quarter of an inch into the box, making it more easily accessible.

The water-loaded animals had a slightly greater total intake on load day than on final adaptation day, 21.9 vs. 19.6 in the 2-foot group, and 21.6 vs. 20.7 in the 6-foot group. Considering that the animals had more time to drink on test day and that the goal-box intake was estimated, these correspondences are so close that we can quite accurately describe the total intake on load day as the amount of water needed to restore fluid balance. By inference we can describe the salt-loaded rats' intake in the same manner and by comparing intakes at the time of criterion with total intake we have the percentage of water needed which had actually been ingested. For the increasing concentrations of load these were: 2-foot: 58, 65, 63, and 31%; and 6-foot: 50,

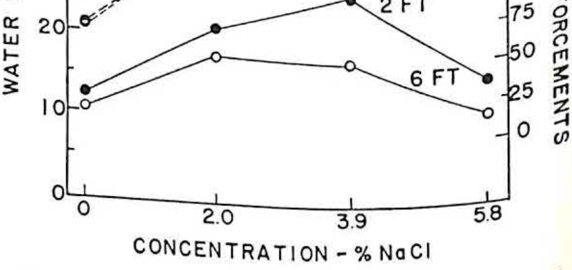


FIG. 19. Load day water intake and number of reinforcements. (The left ordinate is the volume of water consumed, and the right ordinate is the number of reinforcements, from which reinforcement intake is estimated. The upper curves are total intake on test days: load + reinforcement + water consumed during the hour post-session drinking period.)

51, 40, and 22%. The *S*s in the long runway were generally getting about 10% less of their total than were those in the short runway.

Time Scores

During training, the starting-time scores decreased steadily, but the animals in the 2-foot runway consistently started much faster than those in the 6-foot runway. Running times were about the same for both groups, when measured in feet per second. On load day the only differential effects of treatment on running, other than those due to differences in attaining criterion, were the slower starting times of the H_2O and 6% groups as compared with the 2-foot runway 2% and 4% groups in the 2-foot runway (medians of the first five trials). On the day after loading, both starting and running times were significantly slower, but there was almost no differentiation among the variously treated subgroups and the animals in both runways increased about equally. Over the four post-load days these times gradually returned to their previous levels.

Weight and Drinking Changes

As found in the prior drinking experiments, there were increased weights and decreased drinking on the day after loading, but interestingly, the 6% groups had smaller changes than the 4% groups. This difference was significant at the .02 level for weight changes. The reason for this is not clear, but in the repeated load study, on the day after the first load, this occurred also. After the remaining five load days in that study, however, the 6% group did show the greatest change so the effect is still unclear.

Drinking was depressed following load day as a function of both concentration and runway length, rather surprising since both runway groups drank the same total amount on load day. This may, in fact, be no more than sampling error, though significant beyond the .01 level for both variables. The 2-foot group showed the least change.

Gastrointestinal Data

Since these data are the same as those in the clearance study, the reader may refer back to Figures 15–18. As indicated there, the amount of fluid in the stomach and intestine increases with the concentration of the load, and even with the 2% NaCl load the total fluid content of the stomach and intestine at 30-minutes post-loading exceeds the volume of the load. The gut data were correlated with the various running performances, but yielded no significant looking trends within the various subgroups. This lack of relationship may be for any of several reasons, such as the small number of *S*s per subgroup, low variability of scores for GI contents, or lack of reliability from one loading to another.

Discussion

Since the total water intake increased on the test day proportionally with load concentration, but trials to criterion did not, the loads are seen to have both inhibiting and facilitating effects. There are two aspects of the data to be accounted for: the differential proportion of total intake between the 2-foot and 6-foot groups at the time the criterion was reached, and the nonlinear relationship between trials to criterion and load concentration.

Two mechanisms conceivably were operative with respect to the difference between the two runway groups: differential conditioning and/or the differential effort involved in running the two alleys. Since the *S*s in the shorter alley consistently were faster in starting to run during training than were those in the longer alley, there was presumably some kind of differential conditioning such that a smaller amount of thirst reduction might suppress responding in the group with weaker conditioning. In combination with this, the inhibition developed from running the longer runway should tend to weaken the reaction potential for the running response more than would the reactive inhibition developing in the shorter alley.

The nonlinear results for trials to criterion may be explained by the inhibition from the load itself. The increasing amount of fluid in the gut following loading undoubtedly made running uncomfortable in the 6% groups and perhaps somewhat in the 4% groups. The relatively high incidence of running without drinking in the two extreme groups tends to be consistent with Solarz' report (1958) that, after water loading, running persisted longer than drinking. He discusses this finding in terms of differential excitability and suppression of habits by thirst "receptor" and "satiety" cues, drinking being more sensitive to changes in water deficit, increase or decrease, than running.

The present experiment controlled as much as possible the spacing of trials on test day and the delay between running and loading. These variables are interrelated insofar as either of them could allow differential changes in the transfer of fluid within the gastrointestinal tract which might in turn affect running significantly. The importance of the spacing of trials is indicated by comparing three animals loaded, respectively, with water, 2% and 4% NaCl and run in the 2-foot runway with trials separated by 90–120 seconds. Running the animals in rotation, the *E* extinguished after each had run 119 trials (compare with results in Table 2). A 6% animal ran only 16 trials, however, well within the range obtained in the experiment proper. The complex nature of the problem is further indicated by the 6% animal in the 2-foot group who ran 141 trials after loading. Either this *S* was not influenced by what we conceive to be the important inhibitory factors for the 6% group, or his stomach and intestine cleared the load in a very different fashion than did those of his running mates on load day. The amount of fluid in his GI tract following gutting, however, was just about at the mean of the 6% group.

The starting scores in this experiment were indicative of motivational effects induced by our loading, but the highly massed trials precluded precise information. It is probable that behavioral research using this technique of manipulating motivation should employ single trials, or some more optimal spacing. Since only one animal in the present experiment failed completely to run, there should be little attenuation from that source.

GENERAL DISCUSSION

The Linear Scaling of Thirst

The results we have reported on facilitation of drinking by loads of hypertonic saline, when considered together with our findings of drinking inhibition by water loads varying in volume indicate the possibility of arranging load conditions in such a fashion as to produce a linear change in water consumption over a very wide range. If we may assume that drive strength in this situation is directly proportional to consummatory response, we see that stomach loading provides a technique for securing a linear scale of drive strength. It is of interest, therefore, to find a series of loads which will produce increments of the same slope, but opposite sign, as the H_2O decrements. Such a series is shown in Figure 20. The NaCl points were determined by obtaining a least squares fit to the H_2O points, then selecting values from the data of Experiment 1 which showed a reasonable fit to the upward extrapolation of the water-load regression. It turned out that increasing volumes of NaCl having about a 2% concentration potentiated drinking in 23.5-hour deprived rats to the same degree that the same volume of water depressed drinking. The 3, 4, and 5% volume points were obtained in Experiment 1; the 1 and 2% volume points were additional empirical checks.

The value of this particular exercise lies not so much in establishing the fact that drinking can be increased or decreased in a linear manner, since an infinite variety of loads might be used to give the same increments, but rather, that it can be done so easily and with such precision.

From the various experiments we can also summarize the results of loading with a variety of concentrations of 2% body

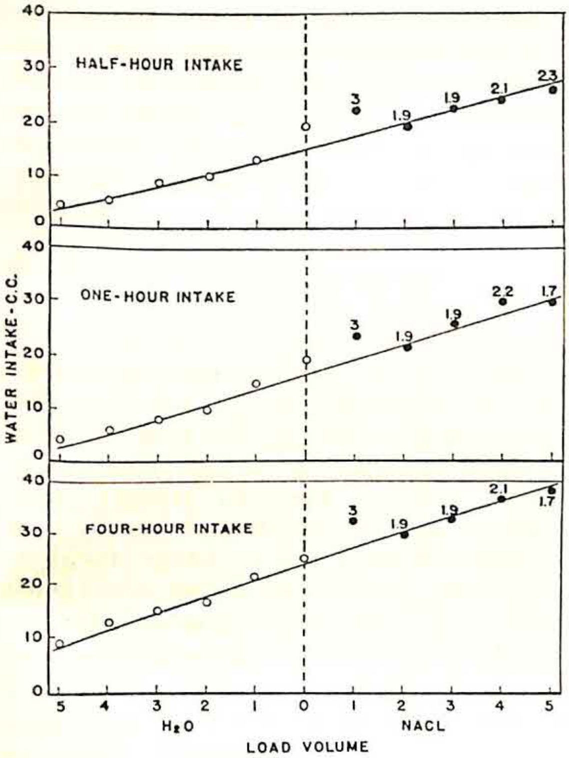


FIG. 20. Linear decrement and increment of drinking in thirsty rats. (The left half of the graph shows intakes at various times after different volumes of water loading—see also Fig. 3—and the right half shows intake after the same volumes of NaCl load. The exact concentrations of NaCl are shown with each point. The vertical dotted line goes through the zero-load points.)

weight volume of load, and compare this with the intake of water as a function of hours of deprivation. This comparison is shown in Figure 21. There is essentially a linear increase in drinking with load concentration up to 5% and then a sharp decline at 6%, on the 30-minute intake. The deprivation curve, on the other hand, is negatively accelerated; 47.5 hours of deprivation leads to the same 30-minute intake as does a load concentration of about 3% NaCl. An extrapolation from the deprivation intake curve to match the highest 30-minute intake after loading would require a number of hours of deprivation far beyond the limits of the rat; the four-hour intakes are totally beyond comprehension in terms of production by deprivation.

We may also refer back to Figure 20, and compare it with the deprivation curve in Figure 21. The highest NaCl load volume in Figure 20 (5%) produces a 30-minute intake which is just about the same as that produced by 47.5 hours of deprivation, but the former has the advantage that equal increments of volume produce equal increments of drinking, rather than progressively smaller increments. In addition, whereas the health of the 47.5-hour deprivation animals steadily declines over 20 days (as indicated by progressive weight loss), the loaded animals remain healthy and appear to thrive on the schedule (as shown, for example, in the weights and daily water intakes over the course of repeated loadings, as shown in Figure 11).

The Effect of Ether

One problem not considered in the experiments proper is the effect of ether per se on water intake, it having previously been shown to depress drinking in a 30-

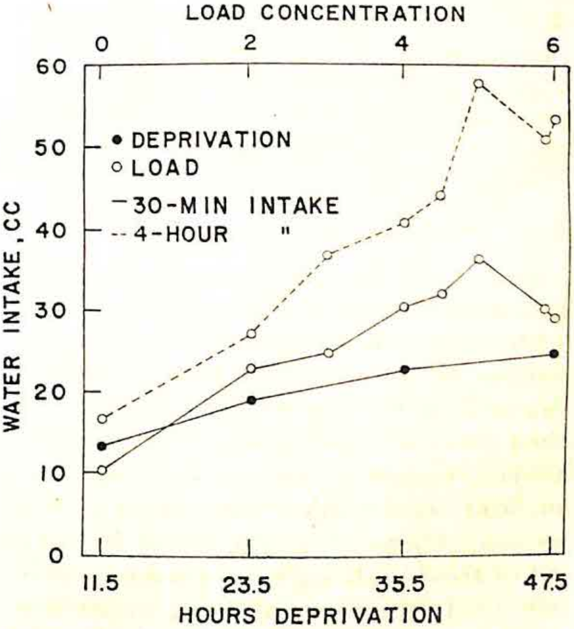


FIG. 21. Comparisons of water intake after varying hours of deprivation (the tenth deprivation interval on each schedule) with intakes following loading with various concentrations of 2% body weight NaCl. (The data for the 11.5-, 35.5-, and 47.5-hr. deprivation points were taken from an experiment run in our laboratory by Jack Dluhy on animals identical in age, sex, and strain to those in the present study.)

minute period (O'Kelly & Weiss, 1955). To estimate the results of etherization in Experiment 1, two additional groups of animals (with *N*s of 6 and 9, respectively) were tested for four-hour intake; one group had 23.5-hour deprivation, the other with deprivation followed by etherization before access to water. These are shown as the "0" load volumes in Figure 20. The etherized group had about 3 cc. less intake in 30 minutes than the nonetherized group, an intake falling directly on the regression line for tap water loads. The nonetherized group, on the other hand, fell above the regression. After four hours, however, these groups had the same intake, indicating that the ether effect had worn off. This is in agreement with the typical findings on humans, that ether is dissipated from the system almost completely in four hours (Goodman & Gilman, 1955, p. 63). Since the load groups and the etherized group all maintain their same relative positions throughout the four hours of drinking, and the no-load etherized group gradually catches up to the nonether group, it would appear that ether has an equally depressing effect on all loads, rather than a differential effect on any particular load.

Load Inhibition

One of the inevitable complications with hypertonic stomach loading (though not with water loads) is that for a period of half an hour or so after loading water is drawn into the stomach and intestine. It then does not clear these compartments for varying lengths of time, which, depending on load concentration, may range as high as several hours (Figures 15 and 16). Not all of these loads appear to produce gastrointestinal distentions inhibiting to performance, however, as witnessed by the fact that in the runway study the 2% and 4% animals did run a great deal more than the water-loaded *S*s. There may have been some inhibition with the 4% load, however, since the animals under that condition ran no more than those with the 2% load, even though they drank more. The 6% animals were clearly inhibited. The most conservative procedure would be to delay the various load groups until the highest concentration group had cleared the load to the same degree as the 2% group in 30 minutes, the delay period used in the runway study (Experiment 6). It is difficult to estimate this exactly since we do not know just *where* in the gut the inhibition occurs. If we look at the intestine (Figure 16) we estimate about 90 minutes; if we look at stomach clearance (Figure 15) we reach an estimate of about two hours for 4% loads and four hours for 6% loads (at which time, according to the delay study, there is little difference between 2% and 4% load groups). Only further research will tell us exactly what the best delay is, but all things considered we would presently estimate about two hours. With this period there is no inhibition of drinking in the 6% load groups, and absorption from intestine is pretty far along. Further experiments are in progress which should tell us more exactly what delays are best.

SUMMARY AND CONCLUSIONS

From all the experiments, the following results stand out:

1. With loading, one can control the water ingestion drinking of thirsty rats over a range from complete inhibition to an intake maximum unobtainable with deprivation.

2. The volume of a hypertonic NaCl stomach load plays a relatively small role in comparison with the amount of sodium chloride in the load. For a given amount of NaCl, the effect of increasing the volume of water is to reduce the intake to the degree that water has been given to the animal in the load itself.

3. Within the range of values studied intensively (0–6%, with 2% body weight volume), there was no harmful effect of the loads over half a dozen repeated loads or with long delays between loading and access to water. There is no reason to believe that this should not hold for all water loads we have used (1–5% body weight) or for all values of NaCl content which fall in this

range (i.e., absolute NaCl contents from 0.3 to 1.2 grams per kilogram of body weight, in solutions between 1–5% body weight).

4. The technique produces highly reliable results over repeated treatments.

5. With repeated loads, water intake tends to increase in a four-hour drinking period but this is offset by an increased urine output and decreased urine concentration such that the load is equilibrated to about the same extent over repeated loads. The increase in intake over repeated loads is the same for all loads used.

6. Following loading, weight increases and drinking decreases as a function of load concentration (and hence, of total intake on load day). A four-day recovery period is probably sufficient, but a six-day period is surely sufficient and is considerably more practical in terms of experimenter convenience.

7. As a function of delay of access to water following loading: (*a*) drinking in four hours decreases, (*b*) post-load weight and drinking changes are attenuated, (*c*) urinary excretion of ions increases, although total urine volume is about the same for all delay groups. The ability of the rat to manage salt loads by renal clearance is an important consideration and should be taken into account in any experiment with loads.

8. With the various hypertonic loads, water is drawn into the stomach and intestine for the first half-hour after loading. The time for clearance and absorption of the load is then an orderly function of load concentration. After loading with high concentration NaCl, drinking and performance are initially inhibited; a delay of one to two hours is suggested when stomach loading of hypertonic NaCl is used in behavioral experiments.

REFERENCES

BRUCE, R. H. A further study of the thirst drive with especial reference to the goal-gradient hypothesis. *J. gen. Psychol.,* 1937, **17,** 49–62.

BRUCE, R. H. The effect of lessening the drive upon performance by white rats in a maze. *J. comp. Psychol.,* 1938, **25,** 225–248.

GAMBLE, J. L. *Chemical anatomy, physiology and pathology of extracellular fluids.* Cambridge, Mass.: Harvard Univer. Press, 1958.

GILMAN, A. The relation between blood osmotic pressure, fluid intake and voluntary water intake. *Amer. J. Physiol.,* 1937, **120,** 323–328.

GOODMAN, L. S., & GILMAN, A. *The pharmacological basis of therapeutics.* (2nd ed.) New York: Macmillan, 1955.

HEYER, A. W., Jr. Studies in motivation and retention: III. A methodological approach to the independent manipulation of tissue dehydration and duration of water deprivation. *Comp. psychol. Monogr.,* 1951, **20,** 251–272.

HOLMES, J. H., & GREGERSEN, M. I. Observations on drinking induced by hypertonic solutions. *Amer. J. Physiol.,* 1950, **162,** 326–337.

MILLER, N. E., & KESSON, MARION L. Is distension of the stomach by a balloon rewarding or punishing? *Amer. Psychologist,* 1954, **9,** 430–431. (Abstract)

MILLER, N. E., SAMPLINER, R. I., & WOODROW, P. Thirst-reducing effects of water by stomach fistula vs. water by mouth measured by both a consummatory and an instrumental response. *J. comp. physiol. Psychol.* 1957, **50,** 1–5.

O'KELLY, L. I. The effects of preloads of water and sodium chloride on voluntary water intake of thirsty rats. *J. comp. physiol. Psychol.,* 1954, **47,** 7–13.

O'KELLY, L. I., & FALK, J. L. Water regulation in the rat: II. The effects of preloads of water and sodium chloride on the bar-pressing performance of thirsty rats. *J. comp. physiol. Psychol.,* 1958, **51,** 22–25.

O'Kelly, L. I., FALK, J. L., & FLINT, DOROTHY. Water regulation in the rat: I. Gastrointestinal exchange rates of water and sodium chloride in thirsty animals. *J. comp. physiol. Psychol.,* 1958, **51,** 16–21.

O'KELLY, L. I., & HEYER, A. W., JR. Studies in motivation and retention: V. The influence of need duration on retention of a maze habit. *Comp. psychol. Monogr.,* 1951, **20,** 287–301.

O'KELLY, L. I., & WEISS, H. H. The Effects of ether and a barbiturate on water regulation in the rat. *J. comp. physiol. Psychol.,* 1955, **48,** 123–125.

PFAFFMAN, C., YOUNG, P. T., DETHEIR, V. G., RICHTER, C. P., & STELLAR, E. The preparation of solutions for research in chemoreception and food acceptance. *J. comp. physiol. Psychol.,* 1954, **47,** 93–96.

SCHALES, O., & SCHALES, S. S. A simple and accurate method for the determination of chloride in biological fluids. *J. biol. Chem.,* 1941, **140,** 879.

SOLARZ, A. K. Effects of hydration on the running and drinking performance of thirsty rats. *J. comp. physiol. Psychol.,* 1958, **51,** 146–152.

TOWBIN, E. Gastric distention as a factor in the satiation of thirst in esophagatimized dogs. *Amer. J. Physiol.,* 1949, **159,** 533–541.

WAYNER, M. J. A study of the effects of sodium chloride and hydration on a monosynaptic reflex pathway in the spinal cord of the albino rat. Unpublished doctoral dissertation, University of Illinois, 1953.

WAYNER, M. J., & REIMANIS, G. Drinking in the rat induced by hypertonic saline. *J. comp. physiol. Psychol.,* 1958, **51,** 11–15.

(Received February 3, 1960)

Vol. 74, No. 14 Whole No. 501, 1960

Psychological Monographs: General and Applied

EVALUATING THE EDUCABILITY OF THE SEVERELY MENTALLY RETARDED CHILD[1]

HELEN SCHUCMAN[2]

Division of Clinical Psychology, Columbia-Presbyterian Medical Center

THE purpose of this study was to investigate a method for obtaining an objective, quantified, predictive measure of educability for children who are severely retarded mentally. The basic hypothesis was that the child's educability can be inferred from his responses to learning situations which require abilities on which education depends, namely, to learn from instruction, to transfer the training, and to retain the learning.

With increasing attention focused on the education of these children, the need for suitable methods of evaluating their abilities has become acute. Their limited abilities must be assessed, and attention directed toward developing the most profitable ones, to whatever extent this is possible.

Unfortunately, it is precisely these children whose abilities are among the most difficult to evaluate by the usual testing procedures. On standard tests, the child may be heavily penalized by verbal, motor, sensory, and experiential handicaps. Further, inability to cope with standard tests is one of the outstanding behavioral characteristics of the mentally retarded. Tests with which a child cannot cope will hardly serve to identify his assets. No curriculum can be built on what a child can*not* do.

It has been said that "the ideal test in the study of mental deficiency would be one which investigated the ability to learn" (Penrose, 1934, p. 49). This research was undertaken to study learning in severely retarded children, and to use the results as the basis for determining what they *can* do.

METHOD

The general method of testing was as follows. A battery of tests was administered to the subjects individually. Each subject was first given the tests without training. He was then trained to the correct responses according to standard training procedures, and retested to permit measurement of posttraining learning gains. Different forms of the tests were used to evaluate transfer. Retention scores were obtained by retesting the subject at various later dates.

Preliminary Experimentation

The final test battery was determined on the basis of the results of a series of pilot studies. The chief aim of the latter was to bring the tests into increasing accord with criteria which were established in advance of the preliminary experimentation. The criteria were as follows:

---

[1] The research reported herein was performed by New York University, School of Education, pursuant to Contract SAE-7783, with the Office of Education, United States Department of Health, Education, and Welfare.

Grateful acknowledgement is made to the following faculty members of New York University: Alonzo Grace, Howard Newburger, Roscoe Brown, Kenneth Benson, Jacob Cohen, and members of the Department of Educational Psychology; to Harold Michal-Smith of the New York Medical College, Christopher De Prospo of the College of the City of New York, Sidney Weinstein of Van Etten and Jacobi Hospitals, Marion Reissenweber of the Institute for the Crippled and Disabled in New York, and William Thetford, Arthur Carr, and Norman Reiss of the Columbia-Presbyterian Medical Center; and to the Mental Retardation Clinic at Flower and Fifth Avenue Hospitals in New York, the National Association for Retarded Children, the New York Association for the Help of Retarded Children, Parents and Friends of Retarded Children of Bridgeport, Connecticut, the Shield of David Institute for Retarded Children in New York, and the Westchester School for Retarded Children.

[2] Formerly at New York University.

1. Tests should be appropriate to the level of ability of the subjects.

2. Tests should arouse and maintain the interest of the subjects.

3. Tests should be suitable for brain-injured subjects.

4. Tests should depent primarily on abilities sufficiently well developed at the age levels of the subjects.

5. Tests should depend minimally on past experience.

Before the pilot studies were undertaken, the following attempts to meet the criteria were made in the construction of the preliminary battery. To aid in meeting the first criterion, tests were designed so that no verbal responses were required, because of the verbal handicaps of the subjects. Instructions were given chiefly in pantomime, verbal directions being limited to simple words, and accompanied by explanatory gestures.

The literature was surveyed for areas suitable to the abilities of the subjects, and imitative abilities, memory, and gross shape, size, and color (brightness) discriminations were chosen. In addition to being adaptable to the subjects' abilities, these areas were thought to be relevant to the special purpose of the tests. Imitative teaching techniques and rote memory tend to be stressed in the training schools, because the children can grasp better what is shown or demonstrated than what is explained (Wallin, 1955, p. 391). On the other hand, simple size, shape, and color discriminations are often fundamental aspects of the curriculum. Further, since almost every child has had some opportunity to develop these discriminative abilities, tests which tap them seem to be among the most favorable measures of differences in ability (Stutsman, 1931, p. 45).

In the search for materials which would arouse the interest of the subjects, toys and training materials chosen by children in the training schools were noted. Sample materials were then constructed, and each sample was included in a group of toys presented to at least 12 severely mentally retarded children. Materials were considered only if they were spontaneously selected by more than half the children to whom they were offered. Short tests were planned, to hold the subjects' interest. The number of items in the individual tests ranged from four to eight, and were arranged in increasing order of difficulty.

Procedures used to meet the criterion of suitability for brain-injured subjects were in line with work done by Cotton (1939), Arnold (1945), Lord (1937), Strauss and Lehtinen (1950), Reissenweber (1953), Strauss and Kephart (1955), and Haeussermann (1958).

Materials were three-dimensional and relatively large. Surfaces were dull to minimize the reflection of light. Tests were not timed. The examiner manipulated the test materials at the subject's direction, if he could not do so himself. Only minimal responses, such as simple choices among the test materials, were required. Materials were presented far enough apart so that a gross motion of hand or arm sufficed to indicate the subject's selections. An achromatic scale was used, because of possible difficulties in color perception. Materials were painted black, white, and various grays in the Munsell Gray Scale, and were presented on a cloth of approximately the middle gray in the scale.

Developmental norms from the literature served as guides to aid in meeting the fourth criterion, and abilities usually developed at least one year below the age of the youngest subjects tested were selected. Two methods were used in the attempt to meet the requirement of minimal dependence on past experience. First, tests were sought which provided the experience necessary for performance in the testing and training procedures themselves. Otherwise, tests were considered if it was reasonable to assume that the materials and procedures were within the experience of all of the subjects.

Transfer situations were constructed in the light of E. L. Thorndike's theory, which holds that transfer depends on identical elements in the original learning and the new learning which it facilitates (Hilgard, 1948, p. 29). The transfer forms were developed in the course of the pilot studies by introducing various changes into the materials and/or the procedures used in the original forms of the tests. In order that the transfer forms would remain within the abilities of the subjects, only relatively few and comparatively simple changes were made. Some of the principles followed in the construction of the transfer forms were based on procedures used by Lashley (1938), Harlow (1949), Barnett and Cantor (1957), and House, Zeaman, Orlando, and Fischer (1957).

The sample used in the preliminary experimentation consisted of 115 mentally retarded children. Of these, 80 were between the ages of 5–0 and 12–6, and were severely retarded according to the criteria used in the selection of the final sample, as stated below. Seven pilot studies, each using 10 subjects, were conducted with this group. The remaining 10 were used for informal experimentation with various changes in materials and procedures, in advance of each pilot study.

In addition, eight severely retarded subjects between the ages of 3–6 and 4–1, and six children aged 12–7 to 14–0 were tested, to aid in determining the most suitable age range for the administration of the final battery. Ten moderately retarded subjects, and 11 whose retardation was attributed primarily to schizophrenia according to their medical and psychiatric records, were also tested on an experimental basis.

The original battery consisted of 11 tests. Preliminary work with some of these tests is reported by Schucman (1957-58). At least two tests in each area were included in the preliminary battery, since it was expected that some of them would prove to be unsatisfactory.

All of the tests were administered in the first four pilot studies. The less suitable ones began to be dropped thereafter. In the earlier pilot studies, only pretraining, posttraining, and transfer scores were obtained. The later ones also included retesting the subjects at various later dates, to aid in determining suitable time intervals at which to measure retention.

Several transfer forms for each test, based on changes in the original materials, in the procedures, or in both, were tried in the preliminary experimentation. The following procedures were used in introducing change into the materials. For tests of shape discrimination, brightness or size was altered, and shape was held constant. In size discrimination tests, shape or brightness was changed, and size kept constant. In test of brightness discrimination, brightness was held constant, and shape or size was changed.

Various alterations in the original procedures for administering the tests were also studied in the preliminary experimentation with the transfer forms. In one method, the original materials were presented, but in a different order. In a second method, the original demonstrations to the subject were included, but without the explanatory gestures which had accompanied them. In a third, the demonstrations were omitted, and only the explanatory gestures were given. In a fourth, the test was administered according to all of the original procedures, but in a different place.

Materials, procedures, and the order in which the tests were to be administered were determined in advance of each pilot study, and a list of criticisms was drawn up afterwards. Tests were changed, substituted, accepted, or rejected on the basis of their adequacy in meeting the test criteria. Quantitative and qualitative analyses of the data were made after each pilot study, to indicate strengths and weaknesses in the tests. The preliminary battery, and its evolution in the course of the pilot studies, are described in Appendix B.

Selection of Subjects for the Final Battery

The sample to which the final test battery was administered was limited to subjects suffering from various pathological physical conditions known to produce mental retardation, as stated in their medical records. Most of these records were obtained from special clinics for the mentally retarded. Some were supplied by physicians on the staff of the school which the child attended. In a few cases, the report of the child's personal physician was used.

Medical diagnoses ranged from relatively clear-cut statements of diagnostic category (such as "mongoloid") to more general, descriptive statements (such as "brain-injured"). Subjects whose retardation was ascribed primarily to "emotional difficulties" or schizophrenia, and categories such as "familial," "undifferentiated," "unclassified," and "etiology undetermined" were excluded from the sample.

In limiting the sample to severely retarded subjects on the basis of psychological test data, only evaluations made within the year, by qualified psychologists attached to the clinic or school which the child attended, were used. An IQ below 50 on the Revised Stanford-Binet Intelligence Scale was used as the criterion for all of the subjects to whom the scale could be meaningfully administered. Sixty children fell into this category. For subjects to whom the Binet scale could not be given, an IQ below 50 on any other standard intelligence test, such as the Merrill-Palmer Scale, the Cattell Intelligence Scale for Infants and Young Children, and the Kuhlmann-Binet Scale, was accepted. This criterion was used for an additional 41 subjects.

An estimated IQ below 50 was used for the 13 subjects who could not, or did not, respond to any of the standard intelligence tests in a way that permitted meaningful scoring. An estimated IQ was a measure made by the staff psychologist of the school or clinic, who had evaluated the child at least once during the past year, and who was personally familiar with him. The psychologist based the estimate on a study of the child's psychological record, which included the results of whatever standard test could be administered to the particular child, analyses and interpretations of the scores obtained, observations of the child's behavior, and clinical impressions of his level of mental functioning. The Vineland Social Maturity Scale, administered to the parent, was the only standard test that could be used for all of the subjects.

The age range for the final sample was set at 5–1 to 11–11. Results obtained in the pilot studies suggested that above that age level the tests tended to become too easy to permit discrimination among the subjects, and below it, the tests tended to be too difficult.

Screening Procedures for Sight and Hearing. The subjects' vision and hearing were evaluated by the following procedures. Medical records, teachers' reports, psychological evaluations, and questionnaires filled in by the parents were studied. Trained personnel familiar with the child were consulted. If these sources gave no indications that the child's sight and hearing were too seriously impaired, the examiner undertook special screening procedures for the specific purposes of the tests.

Before the child was brought into the room, a number of the smaller test materials were placed on the table, covering approximately the area to be used in the testing. If the child picked up these materials, reached toward them, or glanced from one to another, it was assumed that his sight was adequate for test purposes. The child's responses to remarks made by the examiner from various positions behind the child were also noted. Materials involving sound were introduced at several points outside his range of vision. If he turned, reached, or glanced in these directions, it was assumed that his hearing was adequate for the purposes of the tests.

Sample. The sample consisted of 114 mentally retarded children drawn from the Greater New York area, living in the community, and attending special classes and/or clinics for the mentally retarded. Their estimated IQs[3] ranged from 8 to 55, with a mean of 36.18 and a standard deviation of 12.01. They had attended school from 0 to 36 months, mean 11.52, standard deviation 10.84. Their ages ranged from 61 to 143 months, mean 95.16, standard deviation 23.48. Thirty-seven were mongoloid, and 11 were cerebral-palsied. The remainder suffered from various pathological conditions known to produce mental retardation, such as microcephaly, hydrocephaly, cretinism, and congenital cerebral hypogenesis. There were 69 boys and 45 girls.

Final Battery

The five tests which comprised the final battery are briefly described below. More detailed descriptions of materials and testing and training procedures are given in Appendix A.

I. Two-Buttons Test was primarily a test of imitative ability and memory. Materials consisted of a number of buttons, and two transparent boxes into which the buttons were dropped in alternating order. Any gross motion in the direction of the correct box or boxes constituted a correct response on the part of the subject. The test included six items, the buttons being presented to the subject in increasing numbers. Before each item, the examiner demonstrated the correct response. The subject was then requested to imitate the demonstration. In the transfer situation, the demonstrations were omitted, and only the original explanatory gestures were made.

This test was found to be a good beginning for the battery. A few preliminary clicks of the buttons in the boxes usually aroused immediate interest. Giving the child the buttons to play with, in advance of the test, was an excellent way of establishing rapport.

II. Buzz Test was chiefly a test of size discrimination. It consisted of a number of round, black discs of various sizes, attached to a gray board. Separate black discs of comparable sizes were presented to the subject, who was shown, in pretest demonstrations, that placing a disc over the corresponding sample on the board produced a continuous, buzzing sound. The subject was required to indicate the correct placement on the board for each of the discs presented to him. There were six items, arranged in increasing order of difficulty. In the transfer form, white discs, corresponding in sizes to the black ones, were presented to the subject, to be matched to the black samples on the board.

This proved to be a satisfactory second test in the battery. The "buzz" was sufficiently different from the sounds in the preceding test to evoke fresh interest and curiosity. Also, as one of the shorter tests, it was well suited to a period in which most of the subjects were still "settling down."

III. Music Box Test was largely a test of brightness discrimination. Round music boxes of equal size were used, one white, one black, one gray, one black and gray, and one gray and white. The subject was shown, in pretest demonstrations, that only the gray box "made music." Selection of the gray music box, made in any unambiguous way at the disposal of the subject, was the only response required. There were six items. The gray box was placed in a different position for each item, and the other boxes were added in increasing numbers. Larger, rectangular boxes, comparable in other respects to the round ones, were used for the transfer form.

This was the longest test in the battery. It was found to be most satisfactory when given in middle place, at a time when the subjects were usually sufficiently "settled," and when restlessness tended to be low.

---

[3] Estimated IQs up to 55 were used, provided the subject's score on the Stanford-Binet scale was below 50. Estimated IQs tended to be several points higher than Stanford-Binet IQs.

IV. Light Test, primarily a test of shape discrimination, consisted of a gray box with a plastic-covered window in each side. Knobs of various shapes appeared on each of the sides, arranged in increasing numbers. The subject was shown that only the rectangular knobs, one of which appeared on each side of the box, lit up a series of flashlight bulbs placed behind the windows. A correct response required the selection of this knob, made by the subject in any way possible to him. The test included eight items, each of the sides of the box being presented twice. A second gray box, with black knobs, corresponding in shape to the original white ones, constituted the transfer form.

As one of the shorter tests, this was chosen to follow the longest one in the battery. It was perhaps the best-liked of the tests, and was well suited to overcome developing restlessness and to restore flagging interest toward the end of the battery.

V. Three-Buttons Test, a second test of brightness discrimination, involved sorting white, gray, and black buttons. The examiner dropped one sample button into each of the three transparent containers. The subject was then requested to indicate the box which contained the sample corresponding to each of the buttons presented to him. The test contained six items, the buttons being presented in increasing numbers. In the transfer form, the positions of the sample buttons were reversed, and the subject was required to place the buttons according to the new pattern.

This test was placed at the end of the battery largely because a marked tendency toward perseveration resulted unless the two "buttons" tests were widely separated. Of the two, this one required less time to administer, and made a short, interesting final test.

Test Administrations

The final battery was administered nine times. Administrations A and B measured pre- and posttraining performance, respectively. The transfer forms were used in Administration C. These three administrations were completed for each test before the next one was begun. The subject was returned to his teacher or parent for the interval of approximately an hour which elapsed before the next administration was given.

Administration D measured retention without additional reinforcement. Pretest demonstrations and instructions were omitted, training was not repeated, and transfer forms were not used. Intervals of more than an hour for this method of retesting were found in the pilot studies to produce scores which were too low for purposes of discrimination. All of the subjects received the first four administrations of the battery.

Administration E was given approximately one week after the original testing. Training periods and the use of the transfer forms were again omitted. However, pretest demonstrations and instructions were included, since it was found that after an interval of a week the majority of the subjects in the pilot studies no longer remembered what was required without this amount of reinforcement. Of the original sample, 102 subjects were available for Administration E.

Administrations F, G, H, and I were given approximately seven weeks after the first testing period. They were repetitions of Administrations A, B, C, and D, respectively. Training was again included, and both forms of the tests were used. Thirty-six of the original subjects were available for the final four administrations of the battery.

Administrations A, B, and C of the battery, and the comparable administrations in the later retesting, required an average of about 40 minutes to complete. Administrations D and I could usually be given in a quarter of an hour. An average of about 20 minutes was needed for Administration E.

General Testing Procedures

Testing in the pilot studies was done by the investigator. Qualified, advanced graduate student testers administered the final battery. They were given a manual, and were supervised by the investigator in administering the tests with at least 10 subjects. These scores were not included in the results.

Subjects were tested individually, in the school or clinic which they attended. The number of sessions used to complete the battery depended on the needs of the particular child. Rest periods were sometimes necessary during the first three administrations. They were allowed only between the completion of Administrations A, B, and C of a particular test, and the beginning of the next one. If more than one session was necessary, the time intervals between the sessions were kept parallel when the child was retested. In most cases, however, the first three administrations were completed in a single session.

The child's handedness was determined in advance, if possible. In the relatively few cases where this was not possible, right handedness was assumed. Test items were presented toward the child's favored hand, the middle point of the item corresponding roughly to the midline of the shoulder on that side.

Since the examiner manipulated the materials for the child, if necessary, the subjects were given equal opportunities to see the completed responses, and to make or indicate spontaneous corrections. All responses, both in testing and training, were corrected or correctly completed by the examiner, so that the subjects had equal opportunities to see the correct response the same number of times.

General Training Procedures

It was found in the pilot studies that training the children to both the correct responses and to the avoidance of incorrect ones tended to be too confusing. Therefore, for the final battery, the subject was trained to the correct responses only. In order that the majority of the subjects would succeed in the training periods, extremely simple procedures were followed.

In general, training consisted of demonstrations of the correct responses by the examiner; imitation of this demonstration by the subject, with cues supplied by the examiner; and prevention or immediate correction of errors. Most of the subjects were able to achieve a number of correct responses under these conditions. The items were presented for training in the same order in which they were given in the tests. The subject was trained to a maximum of three times on each item, beginning with the first item failed on the test. Exceptions to these procedures are noted in Appendix A.

Scoring Procedures

The subject was given one point for each correct response. Scoring booklets contained one page for each test, on which each administration was scored separately. To minimize possible contamination in scoring, the original booklets were taken from the testers after the completion of Administration E. Also, the testers were not originally informed that any further testing was anticipated at that point. New scoring booklets were used for Administrations F, G, H, and I.

Teachers' Ratings

Ninety-nine of the subjects attended special schools for mentally retarded children. Teachers' ratings were obtained for this group, from the child's current teacher in whose class he had been for a period of at least one month. A five-point rating scale was used, on which the teacher was asked to rate the child's learning ability, as compared with that of the other children who attended that school. The various populations of the different training schools were thought to be sufficiently similar to warrant the use of this frame of reference.

RESULTS[4]

Test-Retest Comparisons

In order to determine the stability of the battery, product-moment correlations were computed between the scores obtained on comparable administrations of each test, by the 36 subjects who were still available for retesting after an interval of seven weeks. The coefficients, all of which were highly significant, are shown in Table 1.

TABLE 1

PRODUCT-MOMENT CORRELATION COEFFICIENTS* OBTAINED BY TEST-RETEST COMPARISONS ($N = 36$)

Administrations	Test I	Test II	Test III	Test IV	Test V
A and F	.66	.82	.85	.79	.80
B and G	.93	.95	.94	.92	.95
C and H	.93	.90	.91	.92	.96
D and I	.95	.90	.95	.92	.95

* All significant at .001 level.

It can be seen that the coefficients produced by the pretraining administrations of the tests (Administrations A and F) were always lower than those obtained between comparable learning scores. The former ranged from .66 to .85 and the latter from .90 to .96, with no overlap between the ranges. The likelihood that the A-F coefficients would be consistently lower than those between the learning scores is exceedingly small ($p < .01$) by the Mann-Whitney U test (Siegel, 1956).

---

[4] Statistical tables, referring to the original investigation and entitled *Supplement to Measurement of the Educability of Severely Mentally Retarded Children* (Schucman, 1959), are available on interlibrary loan from: Gould Memorial Library, New York University; New York 53, New York.

Intertest and Intratest Comparisons

To determine the intertest consistency of the battery, product-moment correlations were performed between the scores from the same administrations of the different tests, i.e., scores from Administration A of each test were compared with scores from Administration A of every other test, and so on. The 102 subjects who remained available for Administration E were used in these analyses, which were based on scores from Administrations A, B, C, D, and E of each test. The correlations thus resulted in 50 coefficients, 10 for each of the administrations.

All of the coefficients reached the .001 level of confidence. However, those yielded by the pretraining scores, which ranged from .31 to .63, were significantly lower than those produced by the 40 comparisons between the learning scores, which ranged from .72 to .87, with no overlap between the ranges ($p < .01$ by the Mann-Whitney U test).

To determine the intratest consistency of the battery, product-moment correlations were also computed between the scores obtained from different administrations within the same tests, i.e., between scores from Administrations A, B, C, D, and E of each test, using 102 subjects. Fifty coefficients, 10 for each of the tests, were thus obtained.

Again, all of the coefficients were significant at the .001 confidence level. However, the 20 coefficients obtained between the pretraining and the various learning scores within each of the tests were consistently lower than those produced by the 30 correlations of the various measures of learning performance. The former ranged from .50 to .89 and the latter from .86 to .97, with overlap between the ranges occurring in only one instance ($p < .01$ by the Mann-Whitney U test).

Comparisons of Scores Obtained on Different Administrations

In order to determine whether the different administrations of the battery had produced significantly different results, a series of subjects × conditions analyses of variance (Edwards, 1950) were performed, comparing pretraining, posttraining, transfer, and retention performance of all of the subjects, as obtained on Administrations A, B, C, and D of each test.

The results demonstrated that highly significant differences were obtained on every test, the Fs ranging from 94.90 to 163.97 ($df = 3, 339$). With a single exception (between posttraining and retention scores on Test III), the ts between the means, which ranged from the .05 to the .001 levels of confidence, showed that significant differences were found between the scores from all of the administrations of all of the tests. The mean scores obtained on the various administrations of each of the tests are shown in Figure 1. (In comparing the mean scores from the different tests, it should be borne in mind that Test IV contained eight items, while the others included six).

A second series of analyses of variance was performed between the scores obtained on comparable administrations by the 36 subjects who received all of the administrations of the battery. The Fs between administrations ranged from 12.16 to 29.44 ($df = 7, 245$), all significant at the .001 level of confidence. The ts between the

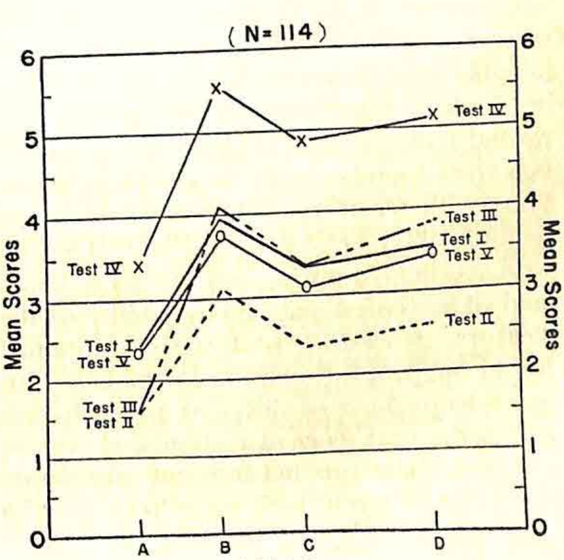


FIG. 1. Differences in mean scores obtained on various administrations.

means obtained on the different administrations demonstrated the following.

Except on Test 1, the scores produced by Administration F were significantly higher than those from Administration A, with *t*s from the .05 to the .001 levels of confidence. Administration F always produced significantly lower scores than Administration B, with highly significant *t*s. In three out of five instances, scores from Administration F were also significantly lower than those from Administration D. It was also found that similar learning curves were obtained across time by Administrations A, B, C, and D on the one hand, and Administrations F, G, H, and I on the other.

Effect of Population Variables on Test Performance

The influence on test performance of five independent variables, namely, sex, time spent in school, age, diagnostic category, and intelligence, was analyzed. Estimated IQs were used as the measure of the subjects' intelligence, since no one standard intelligence test could be validly administered to all of the subjects. Estimated IQs were thought to be the most suitable measure of intelligence for the group for the following reasons.

Estimated IQs, Vineland SQs, teachers' ratings, and Stanford-Binet IQs where obtainable were intercorrelated (r), resulting in highly significant coefficients which ranged from .54 to .92. However, estimated IQs were found to have the closest relationships with all of the other measures, correlating .92 with Stanford-Binet IQs, .82 with teachers' ratings, and .81 with Vineland SQs. To estimate the reliability of the measure, two estimated IQs were obtained for 11 subjects who had been evaluated independently by two different psychologists within the year, once at a clinic and once at a school. The product-moment correlation computed between these estimates yielded a coefficient of .87 ($p < .001$).

Before analyzing the effect of each of the five population variables on test performance, multiple correlations (R) were performed, to determine the extent to which these factors together accounted for the variation in the scores. All of the subjects, and scores from Administrations A, B, C, and D of each test, were used in these analyses.

The coefficients, all highly significant, are shown in Table 2. It can be seen that the coefficients between the pretraining scores and the independent variables, which ranged from .48 to .78, were always lower than those between the learning scores and the independent variables, which ranged from .83 to .90, with no overlap between the ranges ($p < .01$ by the Mann-Whitney U test).

TABLE 2

MULTIPLE CORRELATION COEFFICIENTS* BETWEEN THE FIVE INDEPENDENT VARIABLES AND ADMINISTRATIONS OF EACH TEST ($N = 114$)

Administration	Test I	Test II	Test III	Test IV	Test V
A	.72	.72	.48	.67	.78
B	.88	.87	.86	.86	.87
C	.90	.85	.86	.85	.85
D	.83	.84	.84	.86	.86

* All significant at .001 level.

Expressed as coefficients of determination (R^2), the above results demonstrate that the factors of sex, time in school, age, diagnostic category, and intelligence together accounted for 23 to 61% of the variation in pretraining performance, and 69 to 80% of that obtained in the learning scores.

The influence of each of the five population variables on the scores obtained from the first five administrations of each test was analyzed by a series of "repeated measurements" analyses of variance (Edwards, 1950). Groups were proportionalized so that nonsignificant differences existed between them with regard to each of the four population variables not under consideration at the time. Chi square and t tests were used to test the significance of these differences.

Effect of Sex. Analyses for determining the influence of the factor of sex on the test scores were based on 24 boys and 24 girls. The results demonstrated that this variable did not significantly affect performance on any of the tests, either as a main effect or in interaction with the administrations.

Effect of Time in School. The three levels of time spent in school which were used in the analyses for the effect of this variable on the test scores were based on the length of time of the subjects' school attendance. Nine months were counted for each actual year, the vacation months being excluded.

Subjects with little or no schooling were assigned to the lowest level. There were 28 subjects in this group, who had attended school from 0 to 2 months, with a mean of 1.04. The middle level consisted of 26 subjects, whose school attendance ranged from 3 to 13 months, mean 7.58 months. The 21 subjects in the highest level had attended school for relatively long periods, ranging from 14 to 33 months, mean 21.53.

The results of the analyses demonstrated that the factor of time spent in school did not significantly influence performance on any of the tests, either alone or in interaction with the administrations.

Effect of Age. The levels used for the variable of age corresponded to three general levels of school readiness for this type of population. The groups were somewhat older than is customary with "normal" children at corresponding levels, which is usually the case with severely mentally retarded subjects.

The youngest level corresponded to a preschool group, and consisted of 37 subjects, aged 61 to 83 months, with a mean of 71.00 months. The middle level constituted a beginning and early school group, and contained 36 subjects whose ages ranged from 84 to 112 months, mean 96.39. The 21 subjects in the oldest level were ready, on the basis of age, for more sustained and concentrated training. Their ages ranged from 115 to 143 months, mean 126.81.

Analysis of variance demonstrated that age did not significantly affect performance on Tests I, IV, or V, and no significant interaction effect was found on any of the tests. However, the Fs between the age levels on Tests II and III were both significant at the .05 level of confidence, showing that the factor of age exerted a significant effect on the scores of each of these tests considered as a whole. The between-level ts ($p < .05$) demonstrated that the mean score of the youngest group was significantly lower than that of the middle or oldest levels on both of these tests. No significant differences were obtained between the scores of the middle and oldest groups, or between the age levels on specific administrations.

Effect of Diagnostic Category. Levels used for the variable of diagnostic category were based on the extent of the child's handicaps and the impairment of his general level of functioning. The 55 brain-injured subjects (i.e., children who had suffered trauma to the brain) were assigned to the lowest level. They were thought to be the most handicapped in the sense that various types of physical handicaps were present in addition to the mental retardation. Mongoloids were placed at the highest level, which included 17 subjects. They were considered to be the least handicapped in that, as a group, their motor coordination was relatively good, and their behavior comparatively stable. The middle level consisted of 21 subjects in various diagnostic categories occurring in insufficient numbers to be treated as separate groups, and whose medical records and life histories lacked evidence suggestive of trauma to the brain.

The analyses demonstrated that the factor of diagnostic category did not significantly influence the scores obtained on Tests I, II, III, or V. There was no significant interaction effect on any of the tests. However, an F significant at the .05 level of confidence on Test IV showed that this variable affected performance on the test as a whole. The between-level ts ($p < .05$) demonstrated that the mean score of the mongoloids was significantly higher than those of the other two groups. No significant differences between the levels of diagnostic category were obtained within specific administration.

Effect of Intelligence. Three levels were used for the factor of intelligence, based on the division of the total range of estimated IQs approximately into thirds. The lowest level contained 21 subjects, whose estimated IQs ranged from 8 to 25, mean 15.95. The middle level, which consisted of 31 subjects, ranged in estimated IQs from 27 to 39, mean 34.10. The estimated IQs of the 50 subjects assigned to the highest level ranged from 40 to 55, mean 46.40.

Analysis of variance clearly demonstrated that the factor of intelligence, both alone and in interaction with the administrations, exerted a highly significant influence on the scores of every test. These results are shown in Figure 2, which presents the mean scores obtained by the three levels of intelligence on the five administrations of each of the tests.

The *F*s obtained between the levels of intelligence, all of which were highly significant, ranged from 68.62 to 121.92 ($df = 2, 99$). The between-level *t*s, all significant at the .001 level of confidence, showed that the high intelligence group performed significantly better than the low or middle levels, while the low group was significantly inferior to the middle level on every test considered as a whole.

Further, the *t*s between the intelligence levels on specific administrations, significant at the .001 confidence level in every case, demonstrated that the mean scores of the high intelligence group were significantly greater than those of the low group in all types of learning performance. Also, the low group was always significantly inferior in learning performance to the middle group, and, with one exception, the middle intelligence level was significantly inferior to the high group.

It can therefore be said that significant differences were found between the performance of the three levels of intelligence in 59 of the 60 comparisons based on learning scores. On the other hand, the *t*s for pretraining performance, which were always lower than those obtained for the learning scores, failed to reach statistical significance in 4 of the 15 comparisons.

Of particular interest was the highly significant interaction between the levels of intelligence and the test administrations, which was found on all five tests. The interaction *F*s ranged from 7.01 to 20.95 ($df = 8, 396$). The within-level *t*s showed the following differences in the performance of the different levels on the various administrations of the tests.

The only significant differences found in the scores of the low intelligence level were obtained between pretraining and posttraining scores on four of the five tests. The group failed to transfer the training and

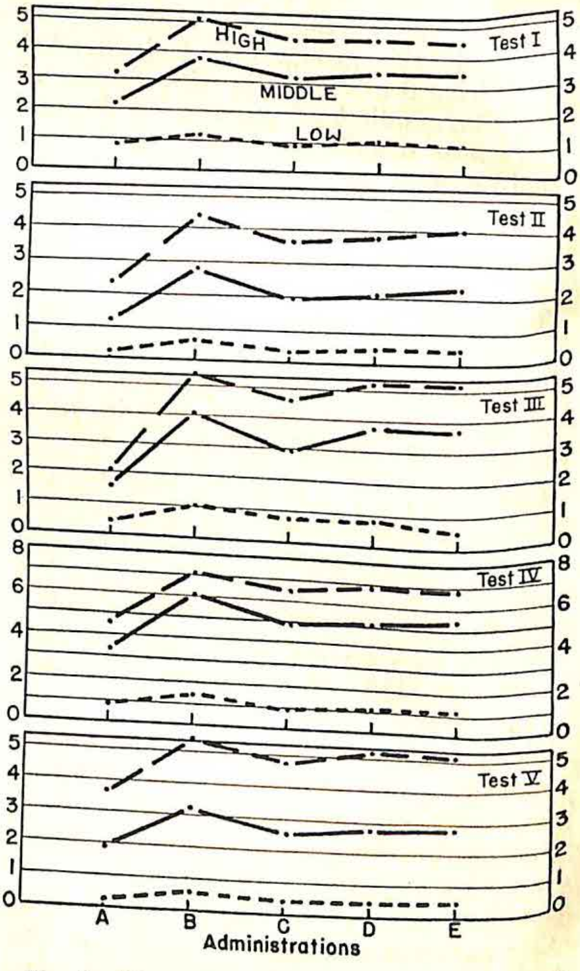


FIG. 2. Mean scores obtained at different levels of intelligence.

to retain the learning to a significant extent on any test.

For the middle intelligence level, all mean scores obtained for learning performance were significantly higher than those from pretraining administrations ($p < .001$). Also, significant differences were found in 17 of the 30 comparisons made between the various learning scores obtained by this group.

Highly significant differences were always found between pre- and posttraining scores for the high level of intelligence. The *t*s between the means were always larger than those for the middle intelligence level, with no overlap between the ranges ($p < .01$ by the Mann-Whitney *U* test). Among the different types of learning scores, highly significant differences were obtained in 21 of the 30 comparisons.

Still another demonstration of the significantly greater influence of the factor of intelligence on learning as opposed to pretraining performance was found in coefficients of determination based on multiple correlations (R^2). All of the subjects, and scores from Administrations A, B, C, and D of each test, were used in these analyses.

The results demonstrated that the variable of intelligence accounted for 18 to 55% of the variation in pretraining performance, and 62 to 76% of the variation obtained in the learning scores, with no overlap between the ranges ($p < .01$ by the Mann-Whitney *U* test).

Comparisons of Test Scores with External Criteria

The intercorrelations of estimated IQs, teachers' ratings, Vineland SQs, and Stanford-Binet IQs have already been mentioned in the section devoted to the relationships of estimated IQs with the other measures. The coefficients were all significant at the .001 level of confidence, demonstrating that the criteria were not unrelated. In addition to those already reported, the correlations produced a coefficient of .61 for Vineland SQs and Stanford-Binet IQs, and teachers' ratings correlated .72 with Vineland SQs and .54 with Stanford-Binet IQs, the latter being the lowest coefficient obtained.

To estimate the reliability of the teachers' ratings, a product-moment correlation was performed between independent ratings made by two teachers who had taught the same child for at least one month during the past year. Two such ratings were obtained for 25 of the subjects, and correlated .85 ($p < .001$).

Product-moment correlations were computed between the scores from Administrations A, B, C, D, and E of each test, and teachers' ratings, Vineland SQs, and psychologists' estimated IQs. Although the close relationship between estimated IQs and test performance was already demonstrated, this measure was again included here because of the added information supplied in this context.

The analyses resulted in 75 coefficients (5 administrations × 5 tests × 3 criteria). Those coefficients ranged from .31 ($p < .01$) to .87, all except the first reaching the .001 level of confidence. With two exceptions, both based on pretraining performance, the coefficients between the test scores and estimated IQs were always higher than those obtained between the test scores and the other criteria. In 18 out of 25 instances, the relationships between teachers' ratings and the test scores were greater than those between Vineland SQs and the test scores.

The correlations of pretraining scores and estimated IQs yielded coefficients ranging from .44 to .77, while those obtained between the learning scores and this criterion ranged from .70 to .87. Within each test, the coefficients based on learning scores were always higher than those obtained for pretraining performance.

To determine the significance of these differences, *t*s were found for the differences between the pairs of correlation coefficients, one of pretraining score with the criterion and the other of learning score with the criterion, as obtained for each test (Guilford, 1956, pp. 193f.). The results showed that the relationship between learning performance and estimated IQs was always significantly greater than that between pretraining performance and this criterion.

The correlations between teachers' ratings and the test scores produced the following results. The coefficients based on pretraining scores ranged from .47 to .72, and those obtained between the learning scores and this criterion ranged from .67 to .78. Again, the latter were always greater than the former within each test. The *t*s between these differences demonstrated that the relationships between the learning scores and teachers' ratings were significantly greater than those between initial test performance and this criterion in 16 of the 20 comparisons.

Similar results were obtained by the correlations of Vineland SQs and the test scores. The coefficients for initial test performance and the criterion ranged from .31 to .70, while those obtained for the learning scores ranged from .64 to .76, with no overlap within the separate tests. The *t*s performed between the coefficients showed that the learning scores were more significantly related to Vineland SQs than were the pretraining scores in 15 instances out of 20.

DISCUSSION

Pilot Studies

Although the immediate purpose of the pilot studies was to increase the adequacy of the tests in meeting the criteria established for them, a number of ancillary findings were obtained from the preliminary experimentation which suggest further research in two main directions. The first is the possible extension of the method of testing used in this study to other groups of children. The following findings are relevant in this connection.

Several tests in the experimental battery, and some of the earlier forms of the final tests (see Appendix B), were found to be too difficult for the sample used in the present study, and were therefore not used in the collection of the final data. However, they were administered experimentally to 10 moderately retarded subjects aged 5–8 to 11–1, with IQs on the Revised Stanford-Binet Intelligence Scale ranging from 51 to 74.

Though the number of subjects was too small to warrant generalizations, the obtained results suggest that these tests are suitable for evaluating learning abilities in moderately retarded subjects, and may provide the basis for an upward extension of the present battery.

On the other hand, some of the tests in the original battery, as well as some of the early forms of the final tests, were found to be unsuitable for the present study because they were too easy for the subjects. These tests were tried with eight severely retarded children, aged 3–6 to 4–11. Although the number of subjects was again very small, the results indicate that these tests can be used for severely retarded children below the age levels used in this study, and may be a basis for a possible age scale.

Of considerable interest are the results obtained by 10 subjects diagnosed as schizophrenic on the basis of medical and psychiatric records. These subjects were among a group of 24 children, who had already been evaluated by a psychologist but whose medical records were not yet available. All of these children were given the final battery, and the results were set aside pending the arrival of the medical records.

Analyses of the scores showed that 11 of the subjects had performed markedly better on all of the tests than did the remainder of the group, whose mental levels were comparable. When the medical reports arrived, it was found that 10 of the 11 subjects were diagnosed as schizophrenic. It is possible that tests such as these may provide a method for differentiating the so-called "pseudoretardate" from the mentally retarded child functioning at a comparable mental level.

Results obtained on tests which were ultimately rejected as unsuitable for brain-injured subjects are also of special interest. Preliminary analysis of the limited results obtained in this area suggest that the brain-injured children were more sharply differentiated from the non-brain-injured subjects by their learning performance than by their initial test scores. It may be that learning behavior, as measured on tests such as these, is a better indicator of the presence of brain-injury than are single administrations of a test or tests.

The second area which warrants further investigation, on the basis of results of the pilot studies, is the possible application of the improvements made in the tests during the preliminary experimentation to curricular materials and training methods in the classroom.

Among the chief difficulties which arose in the development of the battery were the problems of increasing motivation for learning, improving the procedures for training, and minimizing the tendency toward perseveration. A number of successive changes were introduced into the test during the pilot studies, which progressively increased the efficiency of the tests in these respects. The specific set of variables responsible for these improvements cannot be identified at this time. However, the results are suggestive, and may have direct applicability to the training of severely mentally retarded children in the classroom. The findings are as follows:

1. Simple sensory rewards, such as the sound of bells, buzzers, and chimes, and the turning on of lights, were the most satisfactory devices for increasing motivation.

2. These rewards could be built into the materials themselves, or rewards should be supplied by the examiner verbally as the child performed. A combination of both was the most successful.

3. The positive effect of these rewards lasted for relatively short periods. Thereafter, the introduction of another such reward was necessary to reawaken interest and maintain it for another brief period.

4. The factor of novelty appeared to be more important in the rewards than was the particular sense modality involved, i.e., substituting one sound for another was as effective as changing from an auditory to a visual reward.

5. The use of external rewards, such as offers of toys or candy, was often helpful in establishing and maintaining rapport, but was not found to be a significant factor in motivating performance on the tests.

6. The usual testing practices (smiling, nodding, and so on) were highly unsatisfactory. An extremely active and constant participation on the part of the examiner, on the other hand, was a highly facilitating factor.

7. The participation of the examiner was most effective in the form of clapping, laughing, singing, and patting the child.

8. Training to the correct responses only was considerably more productive than training to both selection of the correct responses and avoidance of the incorrect ones. The inclusion of both tended to confuse the subjects.

9. When it was necessary to demonstrate to the subject that certain of the materials were nonfunctioning, it was essential to reinforce the correct response immediately thereafter.

10. The most suitable training methods were those in which all procedures were repeated without variation. Whether this would be so in longer training periods is not known. However, it can be said that the most effective combination for facilitating learning in the short training sessions used for the tests was that of constant changes in the materials, without the introduction of change into the training methods which were used with each of them.

11. Perseveration was minimized when the pattern of the required responses was not changed, and only the number of the responses was increased. The introduction of even minimal changes into the pattern tended to result in an increase in perseveration.

12. Items which required choosing among materials so arranged that the increment in size or degree of brightness between adjacent materials was constant tended to increase perseveration.

13. The tendency to perseverate was minimized when dissimilar objects were placed next to each other, and also when the test items were arranged so that the

selection of adjacent materials was not required.

14. Perseveration was apt to occur when training began with more complex items. Under these circumstances, the subjects were often unable to return successfully to the simpler responses which they had previously given correctly.

15. Single demonstrations of one or two of the simpler items, before undertaking training on the more complex ones, greatly facilitated a later return to the simpler items previously passed.

Adequacy of the Final Battery in Meeting the Test Criteria

The tests were apparently within the abilities of the subjects since, as a group, they obtained significantly higher scores after training, and were able to transfer the training and to retain the learning to a highly significant extent on every test. It can also be said that the tests were able to discriminate among the subjects, since children at different levels of ability produced significantly different scores.

The behavior of the subjects in the test situation clearly pointed to the adequacy of the battery in arousing and maintaining their interest. These qualitative observations were supported by the analyses of the quantitative results. The inter- and intratest comparisons demonstrated a high degree of consistency in the performance of the group, and a definite, consistent, and stable pattern was found in all of the learning curves. Such findings would hardly be likely to occur in the face of lack or loss of interest.

Although the effect of brain-injury on test performance requires further study, the obtained results suggest that the battery was essentially suitable for these children. The brain-injured group succeeded in making significant gains in immediate learning, and was able to transfer and retain the learning to a significant degree. Further, children with no speech and with severe motor handicaps were able to respond adequately.

The obtained results are also thought to show that the tests depended chiefly on abilities sufficiently well developed at the age levels of the group. Although a limited effect of the variable of age was found on two of the tests, significant learning gains were obtained by the different age levels on all of them.

Adequacy in meeting the requirement of minimal dependence on past experience can only be inferred. Perhaps the most directly relevant finding was that the factor of time in school, which ranged from none through five years, did not significantly influence performance on any of the tests.

Differential experiences are necessarily involved in a range of five years of school attendance. Further, school experiences are of a kind which might a priori be believed to be the most likely to influence performance on tests such as these, especially since the materials and procedures were not unlike those used in the training schools. It can therefore be said that precisely those kinds of experiences which might reasonably be expected to exert a differential influence on test performance failed to do so.

In brief, it is thought that the results demonstrated that the battery met the criteria reasonably well. The tests were appropriate to the range of abilities and the age levels studied, succeeded in arousing and maintaining the interest of the children, and were suitable for subjects with severe motor and verbal handicaps.

Effects of Population Variables on Test Performance

No significant influence of the factors of sex or time spent in school was found on the scores obtained from any of the tests. Apparently the battery was suitable for either boys or girls, and for children with school experience ranging from none to five years. The variable of age exerted a limited effect on the scores of some of the tests, its influence being restricted to the performance of the youngest subjects.

No significant influence of the factor of diagnostic category was obtained for the majority of the tests, and only a limited effect of this variable, restricted to the performance of the brain-injured subjects, was found at all. However, these results were to some degree a function of the levels into which the group was divided for the analyses in which the findings were obtained.

It is not thought that the levels of diagnostic category which were used in the present study were adequate for a thorough investigation of its possible effects. Although no rationale for a meaningful subdivision was found with this sample, the use of a brain-injured group as such was open to serious question, since brain-injury is a highly heterogeneous phenomenon. Additional study is also needed in connection with the various categories assigned to the miscellaneous group.

Intelligence was the only factor which was found to be highly related to performance on every test. Of particular interest in this connection are the results which demonstrated that a closer relationship was always obtained between intelligence and learning performance than between intelligence and initial test performance. The factor of intelligence not only exerted a highly differential influence on pretraining as opposed to learning performance, but also on the various types of learning scores obtained, i.e., immediate learning, transfer, and retention.

While subjects characterized by the lowest level of intelligence made significant gains in immediate posttraining learning on most of the tests, they did not succeed in transferring the training or in retaining the learning on any of them. Since the deficiencies of these subjects were most clearly demonstrated in the latter areas, it can be said that transfer and retention scores were the most sensitive of the learning measures to differences in ability.

In contrast to the performance of the subjects at the lowest level of intelligence, the middle intelligence group not only gained in response to training, but was also able to transfer the training and to retain the learning to a highly significant extent on every test. While the high intelligence group also made significant gains in the three learning areas, the magnitude of their mean scores was always higher than that obtained by subjects in the middle intelligence group.

Significant differences in the learning performance of subjects at different levels of intelligence were found in 59 out of 60 comparisons. On the other hand, the magnitude of the differences in their respective mean scores, as obtained by the pretraining administrations of the tests, was always lower than those found in their learning scores. Further, pretraining performance failed to differentiate the groups to a significant extent in 4 out of 15 instances.

These results carry the implication that for severely mentally retarded children, and perhaps for others as well, tests which are administered only once may not be the most sensitive measures of differences in ability.

Reliability

Test-retest comparisons demonstrated a high degree of stability for all of the tests. However, the learning scores on every test were found to be significantly more stable than the pretraining scores, with coefficients of .90 or higher in every case. The learning scores were also found to be significantly more closely related than were the pretraining scores to other learning trials on the same test, and to the same trials on different tests.

These findings suggest that initial test performance was responsive to more specific factors, possibly including previous experience. On the other hand, the greater stability and consistency of the learning scores suggest that a more general factor was measured.

Validity

The obtained results consistently attest to the construct validity of the tests. The evidence on its behalf began with the demonstrated adequacy of the battery in meeting the criteria established for the tests, which was a precondition of validity. In order to measure a child's abilities, he must be placed in situations which are appropriate to his age and to his level of ability, in which he is highly motivated to respond, and in which he is capable of responding within his limitations.

The basis for construct validity lay in the particular learning abilities which were studied. Educability can virtually be defined in terms of a child's ability to learn in response to instruction, to transfer the training, and to retain the learning. Schools are established in order to obtain the desired outcomes of learning, and "it might almost be said that if there is to be education there must be transfer" (Pressey & Robinson, 1944, p. 573).

The particular areas in which the learning abilities were investigated are also thought to contribute to the construct validity of the battery, in that imitative abilities are important in the training of these children, and abilities to make shape, color, and size discriminations are emphasized in the curriculum.

The shape of the learning curves, consistently obtained by the various tests and persisting across time, was also in close accord with the constructs. The lowest scores made by the group were always found in the pretraining administrations of the tests. On the other hand, the highest scores were always obtained from the administration which was given immediately after training, where one would assume that the training would have its maximum effect.

Also in accord with the constructs, the transfer scores made by the group were significantly higher than initial scores and lower than immediate learning gains, indicating that the group successfully transferred some of the training to other situations. Finally, the retention performance of the group was significantly higher than the original scores and lower than post-training learning gains. Although the practice effect which the use of the transfer forms may have introduced is not known, and although the possible influence of each retest on subsequent ones was not determined, the learning was retained to a significant extent, without repetition of the actual training.

The validity of the constructs would also require that the measurements be stable over time, and, if more general learning abilities were tapped, that a high degree of consistency among the learning scores be present. Further, if learning behavior, as obtained on the tests, was related to actual differences in ability, the magnitude of the learning scores would be different at different levels of ability, and would increase as the level of ability increased. It would also be necessary that the learning scores be significantly more sensitive than the pretraining scores to these differences. These results were found on every test.

The results give evidence of predictive as well as of construct validity for the battery. The specific power of the tests to predict the ability of the child to learn in the classroom situation, which would be required if the battery were to be used for purposes of selection or for making actual educational decisions, remains to be demonstrated. Meanwhile, the obtained results offer indirect evidence that the tests would be useful for such purposes. This evidence was obtained in the comparisons of the test scores with the external criteria. Before these relationships are considered, the criteria themselves will be briefly discussed.

Teachers' ratings constituted the closest approach to a criterion of educability per se in this study, in that they reflected the actual experience of the teachers with the children's classroom learning over time. The ratings were made by teachers with special training and experience in the area of mental retardation, who were personally familiar with the children whom they rated. While subjective factors probably entered into their ratings, some amount of objectivity could reasonably be anticipated. The reliability of the ratings was indicated by the highly significant agreement found in the comparison between independent ratings of two teachers evaluating the learning performance of the same child.

A highly significant relationship was also found between the teachers' ratings and psychologists' estimated IQs, each being unaware of the other's estimate. While this agreement may show that the two measures supported each other, it must be borne in mind that the psychologists' records were available to the teachers, so that the extent to which the teachers were influenced by this information cannot be determined.

It can be said, however, that while making the actual ratings, the teachers reviewed their own longitudinal records of the child's performance in the classroom, and did not refer to the psychological records at that time. It can also be said that, although the ratings of the teachers were found to be significantly related to all of the other criteria, the agreement was not so close as to preclude the possibility of independent judgment.

Psychologists' estimated IQs, on the other hand, were found to be very closely related to the scores of standard tests, especially to IQs on the Revised Stanford-Binet Scale where obtainable. This close relationship, of course, may well be due to the fact that the psychologists based their estimates very largely on psychometric data in the children's records.

However, since some amount of subjectivity was probably unavoidable in the estimates of the psychologists, the highly significant relationships between this measure and the scores of standard tests were desirable. The reliability of the measure was indicated by the highly significant correlation between two such estimates made independently by different psychologists evaluating the same child.

The Vineland Social Maturity Scale is a measure of "social competence," which its author defines as "the fuctional ability of the human organism for exercising personal independence and social responsibility" (Doll, 1953, p. 12). The concept is directly relevant to the training of these children, since the curriculum places major emphasis on the development of traits which the construct reflects. This was the only standard test that could be used with all of the subjects, since a social age and a social quotient could be derived on the basis of data supplied by the parent, rather than by the subject himself.

The comparisons of the test scores with the external criteria demonstrated that the results obtained by the various administrations of the tests were significantly related to all of the criteria. However, a marked difference in the degree of the relationship was found for initial test performance on the one hand, and for the learning scores on the other.

In all cases, the correlations between learning performance and psychologists' estimated IQs were significantly greater than those obtained between estimated IQs and initial test scores. In most cases, the relationships found between learning performance and teachers' ratings were significantly greater than those between the pretraining scores and this criterion. Also in most cases, the learning scores were significantly more related to Vineland SQs than was initial performance on the tests.

It can therefore be said that the learning scores obtained were better predictors than were the initial test scores of intelligence, classroom learning, and social competence, as measured by estimated IQs, teachers' ratings, and Vineland SQs, respectively.

SUMMARY AND CONCLUSIONS

The purpose of the study was to investigate a method for obtaining an objective, quantified, predictive measure of educability for the severely mentally retarded child. The method was designed to measure abilities on which education depends, namely, to learn under training, to transfer the training, and to retain the learning.

The subject was first given a battery of tests without training. He was then trained to the correct responses under standard conditions, and retested so that learning gains could be measured. Different forms of the tests were used to study transfer, and retention scores were obtained by retesting the subjects at various later dates.

Criteria established for the adequacy of the tests required that they be appropriate to the abilities and ages of the subjects; capable of arousing and maintaining their interest; and suitable for their verbal, motor, perceptual, and experimental handicaps. Areas selected for study were imitative abilities, memory, and gross shape, brightness, and size discrimination.

A series of pilot studies was conducted with a preliminary battery which consisted of 11 tests. The chief aim of the pilot studies was to bring the tests increasingly into accord with the criteria. Tests were altered, substituted, accepted, or rejected on this basis.

The sample consisted of 114 severely mentally retarded subjects, living in the community, and attending special classes and/or clinics for mentally retarded children. They suffered from various pathological physical conditions known to produce mental retardation. Their estimated IQs (a measure made by the evaluating psychologist, based on quantitative and qualitative analyses of the child's psychological records) ranged from 8 to 55, mean 36.18, standard deviation 12.01. They had attended school from 0 to 36 months, mean 11.52, standard deviation 10.84. Their ages ranged from 61 to 143 months, mean 95.16, standard deviation 23.48. There were 69 boys and 45 girls.

The final battery, which consisted of five tests, was administered to the subjects nine times. The first three administrations yielded pretraining, posttraining, and transfer scores, respectively. The fourth and fifth administrations measured retention, under different conditions and at different times. The last four administrations were repetitions of the first four, given after an interval of seven weeks.

The results demonstrated that the tests were adequate in meeting the criteria established for them. It was also shown that the battery was highly stable. However, as compared with pretraining performance, the learning scores (i.e., posttraining, transfer, and retention) were found to be significantly more stable and significantly more consistent both within and across the tests, suggesting that they measured a more general factor.

Comparisons of the scores on different administrations of the battery demonstrated that the group as a whole learned, transferred the training, and retained the learning to a highly significant extent on all of the tests. It was also shown that the factors of sex and time spent in school did not significantly influence performance on the battery. Age and diagnostic category were found to have a limited effect on the scores of some of the tests.

The variable of intelligence was demonstrated to affect the scores of all of the tests, exerting a highly differential influence both on pretraining as opposed to learning performance, and also on the various kinds of learning performance measured. Transfer and retention scores were found to be the most sensitive measures of differences in ability, suggesting that tests which are administered only once may not be the most discriminating indicators of these differences.

Evidence of construct validity was obtained in the results which showed that learning performance was more stable and more consistent than pretraining performance, in the consistent and stable learning curved obtained by the group on different tests in close accord with the constructs, and in the significantly greater relationship demonstrated between the learning scores and level of ability than between pretraining performance and ability level.

The results also gave partial demonstrations of the predictive validity of the battery. Learning performance, as measured by the tests, was found to be a better predictor than pretraining test performance of intelligence, classroom learning, and social competence, as measured by psychologists' estimated IQs, teachers' ratings, and Vineland SQs.

Within the limitations of the study, the following conclusions are thought to be warranted:

1. The battery was highly stable and consistent.

2. Learning performance was significantly more stable and more consistent than pretraining performance.

3. The factors of sex and time in school did not significantly affect performance on the battery.

4. Age and diagnostic category had a limited effect on the scores of some of the tests.

5. Intelligence exerted a highly significant influence on the scores of all of the tests.

6. Transfer and retention performance were the most sensitive measures of differences in ability.

7. Evidence of construct validity and partial demonstrations of predictive validity for the battery were obtained.

REFERENCES

Arnold, Gwen F. A study of the mental abilities of the cerebral palsied child. Unpublished doctoral dissertation, University of Wisconsin, 1945.

Barnett, C. D., & Cantor, G. N. Discrimination set in defectives. *Amer. J. ment. Defic.*, 1957, **62**, 334–338

Cotton, Carol B. A study of the reactions of spastic children to certain test situations. Unpublished doctoral dissertation, University of Chicago, 1939.

Doll, E. A. *Measurement of social competence.* Philadelphia, Pa.: Educational Test Bureau, 1953.

Edwards, A. L. *Experimental design in psychological research.* New York: Rinehart, 1950.

Guilford, J. P. *Fundamental statistics in psychology and education.* New York: McGraw-Hill, 1956.

Haeussermann, Else. *Developmental potential of preschool children.* New York: Grune & Stratton, 1958.

Harlow, H. F. The formation of learning sets. *Psychol. Rev.*, 1949, **56**, 51–65.

Hilgard, E. R. *Theories of learning.* New York: Appleton-Century-Crofts, 1948

House, Betty, Zeaman, D., Orlando, R., & Fischer, W. *Learning and transfer in mental defectives.* Storrs: Univer. Connecticut, 1957.

Lashley, K. S. The mechanism of vision: XV. Preliminary studies of the rat's capacity for detail vision. *J. gen. Psychol.*, 1938, **18**, 123–193.

Lord, Elizabeth E. *Children handicapped by cerebral palsy.* New York: Commonwealth Fund, 1937.

Penrose, L. S. *Mental defect.* New York: Farrar & Rinehart, 1934.

Pressey, S. L., & Robinson, F. P. *Psychology and the new education.* New York: Harper, 1944.

Reissenweber, Marion. The use of modified block designs in the evaluation and training of the brain-injured. *Psychol. Monogr.*, 1953, **67** (21, Whole No. 371).

Schucman, Helen. A method for measuring educability in severely mentally retarded children: A preliminary study. *Train. Sch. Bull.*, 1957, **54**, 52–54, 58–61; 1958, **55**, 2–4.

Schucman, Helen. *Measurement of the educability of severely mentally retarded children.* Report by New York University to the Office of Education, United States Department of Health, Education, and Welfare, Contract No. SAE-7783, 1959.

Siegel, S. *Nonparametric statistics.* New York: McGraw-Hill, 1956.

Strauss, A. A., & Kephart, N. C. *Psychopathology and education of the brain-injured child.* Vol. II. *Progress in theory and clinic.* New York: Grune & Stratton, 1955.

Strauss, A. A., & Lehtinen, Laura E. *Psychopathology and education of the brain-injured child.* Vol. I. New York: Grune & Stratton, 1950.

Stutsman, Rachel. *Mental measurement of preschool children.* Yonkers: World Book, 1931.

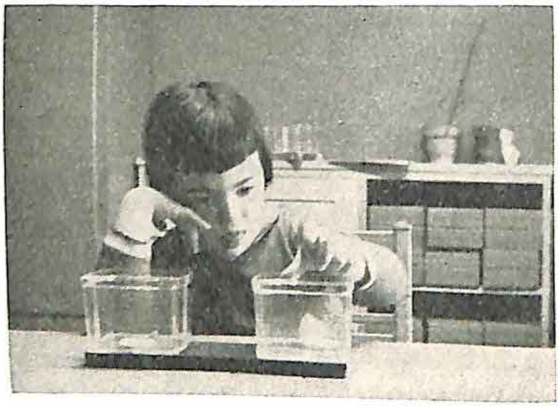
Wallin, J. E. W. *Education of mentally handicapped children.* New York: Harper, 1955.

(Received February 6, 1960)

APPENDIX A

MATERIALS AND PROCEDURES FOR ADMINISTERING THE TESTS[A1]

Two-Buttons Test



Materials. Two clear plastic boxes, 3.5 inches high, with sides 3.5 inches along the bottom, and 4 inches along the top. The boxes were attached, 3.5 inches apart, to a black board 10.5 inches long, and 3.5 inches wide.

Eight round white buttons, 1.75 inches in diameter.

Testing Procedures. There were no special pretest procedures for this test, since demonstrations preceded each item. For each item, the subject was required to indicate the box into which each button presented should be placed, in imitation of the examiner's demonstration. If necessary, the examiner picked up the buttons for the subject, and dropped them into the boxes at his indication.

The test consisted of six items, whch are shown on the scoring sheet (Table A1). Each "L" on the scoring blank represents one button to be dropped into the box at the subject's left. Each "R" stands for one button to be put into the box at his right. The first three items required the correct placement of single buttons, one to be dropped into the box at the left, the next into the box at the right, and the third into the box at the left. The fourth item consisted of two buttons, the fifth of three, and the sixth of four. The buttons were dropped into the boxes in alternating order, always beginning at the left.

Boxes were placed about 2 inches from the edge of the table. Buttons were presented approximately an inch from the edge of the table, and 3 inches apart. Before each item was presented, the examiner demonstrated the correct response, dropping the buttons into the boxes from a height of approximately a foot to ensure a sharp click. As the buttons struck the box, the examiner said, "Ping!" in a singing tone, drawing out the final "ng." After dropping each button, the examiner tapped the box into which it had been dropped, and said, "in *here*." Buttons used in the demonstrations were not removed until the subject's response was secured.

After each demonstration, the examiner placed the number of buttons required for the item before the subject, and said, "*You* do it," pointing from the subject to the buttons, and then to the boxes, midway between them. All buttons were removed before the next demonstration. The examiner said, "Ping!" whenever the subject gave a correct response. If the response was wrong, the examiner corrected it, saying, "Ping!" as the buttons were placed correctly.

The first three items were always given. If all three were failed, the test was discontinued, and training was started. Testing was also discontinued if the subject failed the second, third, and fourth items. If he failed the fourth and fifth items, the sixth was omitted.

Training Procedures. Training began with the subject's first failure unless it had occurred on one of the first three items, in which case training began with the first item. Demonstrations were made as before. The examiner then presented the buttons to the subject, saying "in *here*" for each button, pointing to the correct box, and tapping it until the response was made. Again, "Ping!" accompanied all correct or corrected responses.

The subject was trained three times on each item, unless he achieved two consecutive successes. In that event, the third training was omitted. Training was discontinued if there was complete failure on the first three items. Training was also stopped if the second, third, and fourth items were failed. Training on the sixth item was omitted if no successes were achieved on the fourth and fifth.

The Transfer Form. The original items were those used in the transfer situation. However, the demonstrations were omitted. In advance of each presentation, the examiner pointed to the correct box or boxes, saying, "Ping! in *here*," or "Ping! in *here*" and "Ping! in *here*," and so on. As before, correct or corrected responses were accompanied by, "Ping!"

[A1] Photographs were taken at the Shield of David Institute for Retarded Children.

TABLE A1

SCORE SHEET: TWO-BUTTONS TEST

Administration A

(1) L (2) R (3) L (4) L–R (5) L–R–L (6) L–R–L–R Total________

Training

L	R	L	L–R	L–R–L	L–R–L–R
L	R	L	L–R	L–R–L	L–R–L–R
L	R	L	L–R	L–R–L	L–R–L–R

Administration B

(1) L (2) R (3) L (4) L–R (5) L–R–L (6) L–R–L–R Total________

Administration C

(1) L (2) R (3) L (4) L–R (5) L–R–L (6) L–R–L–R Total________

Administration D

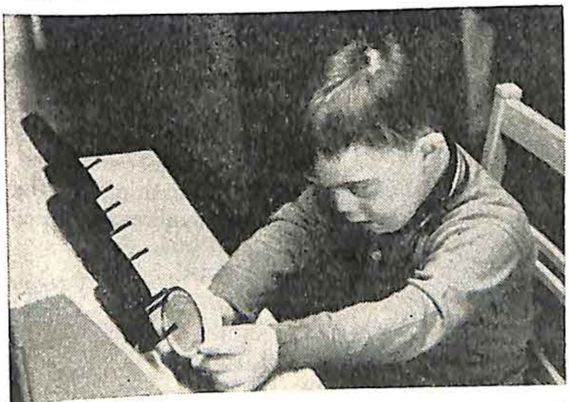
(1) L (2) R (3) L (4) L–R (5) L–R–L (6) L–R–L–R Total________

Administration E

(1) L (2) R (3) L (4) L–R (5) L–R–L (6) L–R–L–R Total________

Comments:

Buzz Test



Materials. A gray board, 23 inches long and 7.5 inches wide, attached at an angle of approximately 45 degrees to a base and a back board, which joined at a 90-degree angle. Four black discs, 1 inch in thickness, and with diameters of 3, 6, 4, and 5 inches, respectively, were attached to the board in that order, with equal distances between them. Thus, beginning at the left, the smallest disc was first, the largest second, the next to the smallest third, and the next to the largest fourth. The sample discs on the board were wired so that, when a disc of corresponding size was placed over one of the samples, a continuous buzzing sound was produced. Two blunt nails projected from the board below each sample, to hold a disc of comparable size in place.

One set of separate black discs and one set of separate white discs, one disc in each set corresponding in size to each of the sample discs on the board.

Pretest Procedures. The board was placed about 4 inches from the edge of the table. The examiner held the smallest black disc at the center of the board, about 1 inch from the edge of the table, at an angle corresponding to the tilt of the board. After holding the disc in this position for several seconds, the examiner placed it over the corresponding sample on the board, saying, "*This* goes *here*." After allowing the resulting buzz to continue for a few seconds, the examiner returned the disc to the center of the board, holding it as before. Pointing from the subject to the disc, and then to the middle of the board, the examiner said "*You* do it." The examiner placed the disc at the subject's indication, if necessary. One demonstration and one unscored response were allowed for each of the black discs. All errors were corrected.

Testing Procedures. For each item, the examiner held a black disc in the position described above, asking each time, "Where does *this* go?" and pointing from the disc to the board. The subject was required to indicate a choice of one of the samples, the examiner placing the disc for him, if necessary. Errors were corrected before continuing to the next item.

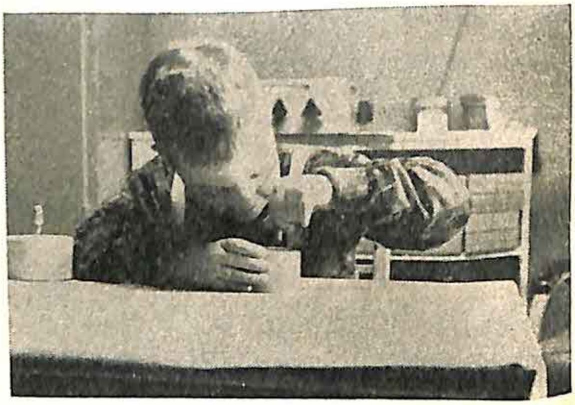
There were six items, each involving matching one black disc to the corresponding sample. The scoring sheet (Table A2) shows the order in which the discs were presented. The number in parenthesis on the scoring sheet refers to the number of the item. The number which follows shows the position on the board of the sample corresponding in size to the disc presented, i.e., "2" represents the largest sample, which was second on the board. All of the items were given, unless the subject failed the first four, in which case the last two were omitted.

Training Procedures. Training always began with the first item. The examiner began by demonstrating the response, saying, "*This* goes *here*," producing the buzz, and then removing the disc and presenting it to the subject, as before. The examiner pointed to the correct sample, and continued to tap it, repeating, "*Here*," until the response was made. If errors still occurred, they were corrected by the examiner.

A special order of the items was used in training on this test, in order to minimize the tendency to perseveration. Instead of the usual procedure of completing training on each item before beginning training on the next, the subject was trained once on all items, in the order in which they were presented in the test, beginning with Item 1. The entire series was repeated three times, unless no errors were made in the course of the first two series.

The Transfer Form. Using the same items, and following the original procedures, the white discs were presented to the subject, to be matched to the black samples of corresponding sizes on the board.

Music Box Test



Materials. Five round music boxes, 2 inches high and 3 inches in diameter. A handle, which could be turned in either direction, projected 2 inches from the center of the box. The shaft of the handle extended vertically from the box for a distance of 1 inch, and was then bent at a right angle, continuing horizontally for another inch, and ending in an upright wooden cylinder, 0.5 inch in diameter and 1 inch long. One box was painted black, one white, one gray, one black and gray, and one gray and white. Only the gray box functioned, playing a repeated series of four tones.

Five rectangular music boxes, 2.5 inches high, 4.5 inches long, and 2.75 inches wide, in other respects like the round boxes.

Pretest Procedures. The round gray box was placed before the subject, about an inch from the edge of the table, with the handle toward him. The examiner turned the handle for some 3 seconds, saying, "music, music, music." The examiner then said to the subject, "*You* make *music*," pointing from the subject to the gray music box. If necessary, the examiner placed the child's hand over or against the handle and turned it for him, repeating, "music, music, music."

TABLE A2

Score Sheet: Buzz Test

Administration A

(1) 1 (2) 3 (3) 2 (4) 4 (5) 1 (6) 4 Total________

....

Training

1 3 2 4 1 4 —— 1 3 2 4 1 4 —— 1 3 2 4 1 4

....

Administration B

(1) 1 (2) 3 (3) 2 (4) 4 (5) 1 (6) 4 Total________

....

Administration C

(1) 1 (2) 3 (3) 2 (4) 4 (5) 1 (6) 4 Total________

....

....

Administration D

(1) 1 (2) 3 (3) 2 (4) 4 (5) 1 (6) 4 Total________

....

Administration E

(1) 1 (2) 3 (3) 2 (4) 4 (5) 1 (6) 4 Total________

....

Comments:

The gray box was left on the table, but was pushed aside, some 6 inches to the subject's left. The round black and gray box was presented in center position, its black side toward the gray box. The examiner turned the handle about 3 seconds, shaking his head and saying, "no music, no, no music." The subject was then asked to turn the handle, as before, the examiner repeating, "no music" as he responded.

The black and gray box was pushed to the subject's right, and the round gray and white box was presented, with the white side toward the gray box. The demonstration and response were repeated. The box was then pushed to the subject's left, and the gray box was presented again. The examiner played it, again for about 3 seconds, saying "*Here's* the music," and asking the child to play it, as before. The boxes were removed before the test began.

Testing Procedures. Boxes were presented about 4 inches apart, the handles turned toward the subject. At each presentation, the examiner said, "*You* make music," pointing from the child to the music boxes. The subject was required to indicate his choice, the examiner turning the handle of the box he selected, if necessary. If the subject's response was wrong, the examiner pointed to the gray box, turned the handle, and said, "*Here's* the music."

The shifts in the positions of the boxes were made with the subject watching. However, to prevent uninterrupted visual fixation on the gray box, the subject's attention was briefly diverted during each shift. To keep the interruption as standard as possible, the examiner said, "Look at me, (subject's name)," and then redirected the child's attention to the table. This required 2 or 3 seconds.

The scoring sheet (Table A3) shows the

TABLE A3

Score Sheet: Music Box Test

Administration A

(1) G–B (2) W–G (3) G–W–B (4) B–W–G–W/G (5) W/G–G–W–B–B/G
(6) W–B–B/G–G–W/G Total________

……

Training

G–B	W–G	G–W–B	B–W–G–W/G	W/G–G–W–B–B/G	W–B–B/G–G–W/G
G–B	W–G	G–W–B	B–W–G–W/G	W/G–G–W–B–B/G	W–B–B/G–G–W/G
G–B	W–G	G–W–B	B–W–G–W/G	W/G–G–W–B–B/G	W–B–B/G–G–W/G

……

Administration B

(1) G–B (2) W–G (3) G–W–B (4) B–W–G–W/G (5) W/G–G–W–B–B/G
(6) W–B–B/G–G–W/G Total________

……

Administration C

(1) G–B (2) W–G (3) G–W–B (4) B–W–G–W/G (5) W/G–G–W–B–B/G
(6) W–B–B/G–G–W/G Total________

……

……

Administration D

(1) G–B (2) W–G (3) G–W–B (4) B–W–G–W/G (5) W/G–G–W–B–B/G
(6) W–B–B/G–G–W/G Total________

……

Administration E

(1) G–B (2) W–G (3) G–W–B (4) B–W–G–W/G (5) W/G–G–W–B–B/G
(6) W–B–B/G–G–W/G Total________

……

Comments:

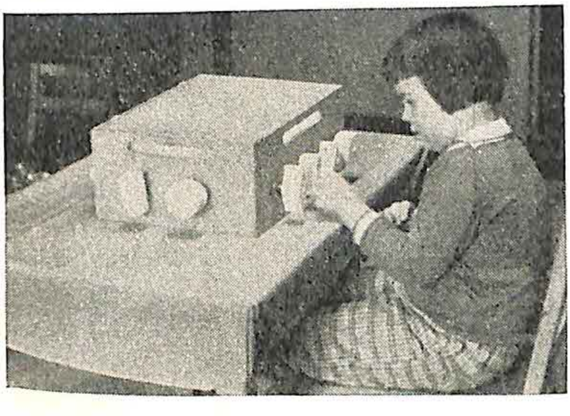
number of boxes used for each of the six items, and the order in which they were presented. The letter "G" on the sheet represents the gray box, "B" the black, "W" the white, "B/G" the black and gray, and "W/G" the white and gray. Two boxes were used for each of the first two items. Three were used for the third, and four for the fourth. Five boxes were used for the fifth and sixth items. The two-toned boxes were not introduced into the test until the fourth item. One of them was used in the fourth item, and both in the fifth and sixth. In the fifth item only one of the two-toned boxes was placed next to the gray box. In the sixth item, the gray box was presented with a two-toned box on each side. The two-toned boxes were always presented with the gray side away from the gray box. The first three items were always given. The test was discontinued if these three items, or the second, third, and fourth items were failed. The sixth item was not given if the subject failed the fourth and fifth.

Training Procedures. If the subject's first failure occurred on one of the first two items, training began with the first. Otherwise, it began at his first failure. The examiner presented the item, turned the handle of the gray box, and said, "music, music, music." Then, pointing from the subject to the gray box, the examiner said, "*You* make *music*," tapping the gray box, and repeating, "*Here's* the music," until the response was secured. If necessary, the examiner placed the handle of the gray box into or against the subject's hand, turning the handle for him, and saying, "Music, *here's* the music." No interruptions were made in the shifts of materials during training.

The subject was trained three times on each item, or through two consecutive successes. If all test items were failed originally, training was always given on the first three items. It was discontinued if these were failed completely, or if complete failure occurred in training on the second, third, and fourth items. Training was omitted for the sixth item, if no successes were achieved on the fourth and fifth.

The Transfer Form. The rectangular music boxes were used in place of the round ones. The original testing procedures were followed, and the original test items were used. The round set was pushed to the back of the testing table, but was not removed, in an attempt to emphasize the fact that different boxes were being presented.

Light Test



Materials. Two gray boxes, each 14.5 inches long, 11 inches wide, and 6.5 inches high. White knobs were used on one of the boxes, and corresponding black knobs on the other. Above the knobs, on each side of the box, was a plastic covered "window," 6 inches long and 1 inch wide. Behind the windows were a series of flashlight bulbs, which lit when the rectangular knob was pressed. A single rectangle appeared on each side. A rectangle and a sphere were used on Side 1. Side 2 contained a diamond and a rectangle. Side 3 included a rectangle, a diamond, and a triangle. A diamond, a rectangle, and a triangle, and a sphere were used on Side 4. The spheres were 2 inches in diameter. The triangles were 2.5 inches on each side. The diamonds were 2.75 inches long, and 1.5 inches wide. The rectangles were 2.5 inches high and 1 inch wide.

Pretest Procedures. The box with white knobs was placed about 2 inches from the edge of the table, Side 1 facing the subject. The examiner pressed the rectangular knob, pointed from the knob to the lighted window, and said, "*Light*, see the *light*." Then the examiner pressed the sphere, pointed from the knob to the unlit window, shook his head, and said, "No light, see, no light." Next, the examiner pressed the rectangle again, pointing as before, and saying, "*Here's* the light." Turning to the subject, the examiner said, "*You* make *light*," pointing to the subject, and then to the box. If necessary, the examiner pressed the knob at the subject's indication, and corrected the response in case of error. As the rectangle was pressed, the examiner pointed again to the window, repeating, "See the *light*." The examiner then turned the box so that Side 2 faced the subject, and the entire demonstration was repeated.

Testing Procedures. The test contained eight items, each consisting of the presentation of one side of the box. The order in which the sides were presented is shown on the scoring sheet (Table A4). As each side was presented, the examiner said, "*You* make *light*," pointing from the subject to the box. If necessary, the examiner pressed the knob which the subject indicated as his choice, and corrected any errors made. All items were given unless the subject failed the first four.

Training Procedures. Training began with Side 1. The examiner pressed the rectangular knob, saying, "*Here's* the *light*," pointing from the rectangle to the lighted window. The subject was then asked "to make light," the examiner pointing to the rectangular knob and tapping it until the subject's response was made. The examiner repeated "*Here's* the light" as the rectangular knob was pressed, and corrected errors if necessary.

The subject was trained three times, or through two consecutive successes, on each side of the box. Training was given on each of the four sides, unless the subject achieved no successes in training on the first three sides.

The Transfer Form. The original procedures were followed, using the box with black

TABLE A4

Score Sheet: Light Test

Administration A

(1) 1 (2) 2 (3) 3 (4) 4 (5) 1 (6) 2 (7) 3 (8) 4 Total________

Training

Side 1	Side 2	Side 3	Side 4
Side 1	Side 2	Side 3	Side 4
Side 1	Side 2	Side 3	Side 4

Administration B

(1) 1 (2) 2 (3) 3 (4) 4 (5) 1 (6) 2 (7) 3 (8) 4 Total________

Administration C

(1) 1 (2) 2 (3) 3 (4) 4 (5) 1 (6) 2 (7) 3 (8) 4 Total________

Administration D

(1) 1 (2) 2 (3) 3 (4) 4 (5) 1 (6) 2 (7) 3 (8) 4 Total________

Administration E

(1) 1 (2) 2 (3) 3 (4) 4 (5) 1 (6) 2 (7) 3 (8) 4 Total________

Comments:

knobs. The first box was pushed to the back of the testing table, but was not removed from sight.

Three-Buttons Test

Materials. Three clear plastic boxes, 3.5 inches high, 3.5 inches along the bottom, and 4 inches along the top. The boxes were placed 3 inches apart, and attached on parallel sides to two wooden runners, 17 inches long, 1 inch wide, and 1 inch high.

Nine round buttons, 1.5 inches in diameter, 3 black, 3 white, and 3 gray.

Pretest Procedures. The boxes were placed about 2 inches from the edge of the table. The examiner dropped a sample white button into the box at the subject's left, saying, "Ping!" as the button struck the bottom of the box. Leaving the sample in place, the examiner placed another white button before the subject, about 1 inch from the edge of the table and approximately at the center point of the middle box, and said, "*You* do it," pointing from the subject to the button, and then to the box containing the sample. If necessary, the examiner placed the button at the subject's direction, and corrected or correctly completed the response. These procedures were repeated with black and gray buttons, using the middle and the right hand boxes, respectively. The examiner added "Ping!" as each button was correctly placed. The sample buttons remained in the boxes throughout testing and training.

Testing Procedures. Buttons were presented about an inch from the edge of the table. Single buttons were presented approximately at the center of the boxes. For the two-button items, one button was placed at the mid-point of the space between the left and center boxes, and the other at the corresponding point between the right and center boxes. For the final item, the middle button was placed at the center of the boxes, and the others at the midpoints of the spaces between the center box and those to the left and right, respectively.

The buttons used for each item were removed after its completion. Buttons used in each of the six items are shown on the scoring sheet (Table A5). "W" represents a white button, "B" a black one, and "G" a gray one. The first three items consisted of placing a single white, black, and gray button into the box containing the corresponding sample, the examiner dropping the buttons into the boxes at the subject's direction, if necessary. The fourth item consisted of a white and a gray button, and the fifth of a gray button and a white one, presented in that order. A black, a gray, and a white button were used for the sixth item. A correct response for the latter

TABLE A5

SCORE SHEET: THREE-BUTTONS TEST

Administration A

(1) W (2) B (3) G (4) W–G (5) G–W (6) B–G–W Total________

Training

W	B	G	W–G	G–W	B–G–W
W	B	G	W–G	G–W	B–G–W
W	B	G	W–G	G–W	B–G–W

Administration B

(1) W (2) B (3) G (4) W–G (5) G–W (6) B–G–W Total________

Administration C

(1) G (2) W (3) B (4) G–B (5) B–G (6) W–B–G Total______

Administration D

(1) W (2) B (3) G (4) W–G (5) G–W (6) B–G–W Total________

Administration E

(1) W (2) B (3) G (4) W–G (5) G–W (6) B–G–W Total________

Comments:

items required the correct placement of all of the buttons presented.

As each item was presented, the examiner asked, "Where does *this* go?" or, "Where do *these* go?" pointing from the buttons to the boxes. The examiner accompanied each correct placement with a "Ping!" Errors were corrected, or correctly completed by the examiner. The first three items were always presented. Testing was discontinued if the subject failed the first three items, or the second, third, and fourth. The last item was omitted if the fourth and fifth were failed.

Training Procedures. Training was given three times on each item, or through two consecutive successes. Training always began with the first item. However, if the subject had passed the earlier items on the test, he was trained only once on those items, regular training beginning at the point of his first failure.

Before each item, the examiner demonstrated the correct placement for each button, presented the item, and pointed to the correct box until the subject had responded, saying, "in *here,*" while tapping the edge of the box. All errors were corrected, and a "Ping!" accompanied a correct response made by the subject, or by the examiner at the subject's indication. Training was discontinued if there was complete failure on the first three items, or on the second, third, and fourth. Training was omitted on the sixth item if the subject failed the fourth and fifth.

The Transfer Form. The samples for the original form of the test were removed, and those used for the transfer situation were placed into the boxes by the examiner, following the procedures used in placing the original samples. The positions of the samples, and the buttons used for each item, are shown in Table A5 (see Administration C). The original testing procedures were followed.

APPENDIX B

The Experimental Test Battery

This section describes the preliminary battery used in the pilot studies. The various revisions of the tests which were ultimately included in the final battery are described first. The rejected tests are described thereafter.

Revisions of Tests Used in the Final Battery

The *Two-Buttons Test* underwent eight revisions in the pilot studies. The test originally contained five items. The first two consisted of single buttons, one to be placed in the box to the subject's left, and the other in the box to his right. Two, three, and four buttons were then presented, to be dropped into alternate boxes. The resulting scores were too low. The earlier pilot studies therefore concentrated on adding various easy items.

The addition of two single-button items was attempted first. This, however, produced a marked increase in perseveration. When the new items required reversing the order of placement used for the original items, the subject found the shift almost impossible to make, and training merely tended to increase this difficulty rather than to overcome it. On the other hand, when the added items duplicated the original order, the subjects continued using that order regardless of the number of buttons presented. Perseveration decreased to a satisfactory extent when the number of single-button items was reduced to three, and only an alternating order was required for all placements.

Results of various changes in the multiple-button items were fairly consistent. Items containing odd numbers of buttons were found to be more difficult than even-numbered items, to the extent that training for a three-button item tended to be less successful than training for an eight-button item. However, once a subject had grasped the concept of placing buttons in alternating order, increasing the buttons in even numbers presented little additional difficulty. Items consisting of six and eight buttons were therefore dropped as superfluous.

Of the various odd-numbered items attempted, only one three-button item was retained, since inclusion of two items using odd numbers of buttons made the test too difficult. The three-button item was tried out in various positions, and was found to be most useful in next-to-last position.

Various transfer forms for this test were also studied. In some, the examiner used the original white buttons for the demonstrations, presenting black and/or gray buttons to the subject. These were found to be too easy, since higher scores than on the retest administration were obtained. On the other hand, when black and white buttons were both presented to the subject, the resulting scores were too low. Omission of pretest demonstrations and instructions also made the form too difficult, the subjects apparently failing to understand what was required of them. However, omitting the demonstrations, but pointing to the correct box or boxes in advance of the presentations, seemed to provide sufficient cues.

Of the various sounds tried out as the verbal accompaniment of the demonstrations, "Ping!" was found to be the most interesting to the children, especially when pitched high and given a drawn-out "ng" at the end. Virtually all of the subjects smiled or laughed at this sound. Because of its evident appeal, the sound was reserved as an accompaniment for correct responses only, to increase motivation and reinforce learning.

The *Buzz Test* was revised six times in the preliminary experimentation. As it was originally constructed, the sample discs were arranged on the board in order of size. A round gray peg projected at right angles from the center of each sample, and extending 2 inches beyond it. Each of the discs presented to the subject had a round hole, 1.5 inches in diameter, at the center. The discs slipped over the pegs projecting from the samples, the buzz being produced by pressing a disc against its corresponding sample on the board. The majority of the subjects were unable to make the necessary size discriminations for this version of the test. It appeared to be especially difficult for the brain-injured children.

The pegs were removed in the next pilot study, and discs without holes were used. To produce the buzz, the presented disc was laid over the corresponding sample, supported by blunt nails driven into the board just below the samples. This change made the test easier, and did not appear to penalize the brain-injured subjects. However, perseveration, especially after training, was so great that some of the subjects did not shift from the first response for the remainder of the test. It is interesting to note that perseveration occurred to a significant extent only on this test and the preceding one. It was virtually absent on all other tests throughout their revisions.

In an attempt to overcome perseveration, the differences between the sizes of the discs were increasd. This was unsuccessful. In the next attempt, the arrangement of the samples was changed, to avoid a regular increase in size. The smallest disc was placed first and followed by the largest. The next to the smallest came next, with the next to the largest, last. Perseveration was almost entirely overcome. However, the test was now too easy.

To increase the difficulty, presenting each disc twice was attempted. Perseveration increased sharply. Presenting the discs in the order in which they were used in the test, and then repeating the series again in that order, overcame the tendency to perseveration, but made the test too long to hold interest. Two items were therefore dropped. One complete series was given, plus two additional items consisting of one placement at the extreme left and one at the extreme right. This did not give rise to perseveration, and shortened the test sufficiently.

The *Music Box Test* was changed six times. In its original form, it consisted of two gray boxes and two white boxes. The one functioning box was black. Since this form was found to be too easy for purposes of discrimination, two additional boxes were used, one gray and one white. However, this increase in the length of the test tended to make the subjects restless. The number of boxes was reduced again to five, and one of the gray boxes was darkened considerably. No marked increase in difficulty resulted. However, when the boxes were repainted so that the darker gray box played, the subjects found it too difficult to distinguish between the two shades of gray.

A two-toned box, half white and half black, was introduced in the next attempt, the single gray box playing. Again the test was found to be too easy. Two two-toned boxes, one black and gray and one gray and white were tried next, presented with their gray sides facing the gray box. This tended to confuse the subjects, especially the brain-injured children. In the final form, the two-toned boxes were always presented with their gray sides away from the gray box, which considerably lessened the confusion, and also made the test more suitable for the brain-injured subjects. Since the original rectangular boxes were satisfactory for the transfer form, no changes were made in their size and shape. However, they were painted to correspond with the round boxes in each revision.

The *Light Test* underwent seven revisions. Only three shapes were used in its original construction: spheres, triangles, and rectangles, the latter being the correct response. Three knobs were the maximum used on any one side of the box. Although this form was too easy for the group as a whole, the first two sides were not changed, the simpler items being necessary for discrimination among the more retarded subjects. The revisions therefore concentrated on increasing the difficulty of the remaining items.

A larger box was constructed. The fourth side now containing four knobs, including a second sphere. Since this did not increase the difficulty sufficiently, all spheres were removed from the fourth side, and a diamond was introduced. This was still too easy. One sphere was therefore added, the rectangle being inserted between the sphere and the diamond. Since this also was unsuccessful, the rectangle was placed between the diamond and the triangle, a change which greatly increased the difficulty of the item. The original sphere on the third side of the box was also replaced by a diamond, which increased the difficulty of this item as well.

The Light Test was so interesting to the subjects that the length of the test could be increased without much risk of loss of interest. Two methods were tried for adding to the number of the items. In the first method, each side of the box was presented twice, before the next side was turned to the subject. This produced highly uneven scores, the second item on each side being considerably easier than the first. In the final revision, the sides were presented in the order of difficulty, the entire procedure being repeated twice.

The *Three-Button Test* was revised four times in the course of the pilot studies. In its original form, in which only single buttons were used, it was found to be too easy for purposes of discrimination. Introducing two-button items did not increase the difficulty sufficiently. On the other hand, proceeding directly from single buttons to three-button items made the test too difficult. In the final form, three single-button presentations were used, followed by two two-button items, with one three-button item as the final presentation.

Rejected Tests

The *Sounds Test,* a test of shape discrimination, consisted of a gray board like that used in the Buzz Test, but 2 inches longer. Projecting from the board were four knobs: an octagon, a triangle, a gear, and a star. Pressing the octagon caused chimes to sound, the triangle rang a bell, the gear produced a buzz. Pressing the star caused no sound. The subject was shown, in pretest demonstrations, the results of pressing the various knobs. He was then asked to reproduce the sound, being instructed "to make ting-a-ling," "to make buzz," and so on. For the star, he was asked "to make sh-sh," the examiner putting a finger to his lips to indicate silence. Various patterns of sounds were used during the pilot studies.

A number of special difficulties arose. One problem was the unwillingness of the subjects to shift from one knob to another. This was apparently not due to a perseverative tendency, since the subjects seemed to be able to shift when they elected to do so. However, they liked the sounds, and, having produced one, they tended to repeat it with evident enjoyment, regardless of instructions.

In an effort to force the shift, cut-off buttons were installed, so that the examiner could shut off the sound after the response was secured. This tended both to confuse and to anger the subjects, who often pressed harder and harder on the knob, apparently to "make it work." In the next attempt, the examiner demonstrated the next response immediately after completion of the previous one, thus directing the child's attention to the new sound. This was successful in facilitating shifts, but reduced the test to a simple imitation problem.

All attempts to secure responses based on sound patterns were apparently beyond the abilities of the subjects, even when only two sounds were used at a time. The test was finally abandoned as too difficult for purposes of discrimination for this population. It was tried out experimentally with a group of moderately retarded subjects, for whom it appeared to be highly suitable. The subjects enjoyed the test, and were able to handle simple sound patterns. A transfer form based on reversals of the original patterns also appeared to be feasible. Since the test can easily be used for presenting patterns ranging from the very simple to the highly complex, it might serve in an upward extension of this method of testing.

The *Rattle Test* was a test of imitation, in which subject and examiner used comparable sets of material. The subject was required to select the correct material, and use it in imitation of the examiner's demonstration. The materials included a rattle, a small pillow, a hairbrush, and a cup. After each demonstration, the subject was asked to "make rattle, rattle, rattle," "to go to sleep," and so on. A correct response required that the subject indicate the material to be used, and the part of the body with which its use was associated, i.e., the head for the pillow, the mouth for the cup, and so on.

This test was so appealing to the subjects that it was retained in various forms until the last pilot study. Unfortunately, the revisions did not succeed in making it sufficiently useful for purposes of discrimination. When the materials were presented singly or in groups of twos or threes, the test was too easy. When more complicated patterns were attempted, the test became too difficult. Similar difficulties were found in the various transfer forms. Reversals of the order of the materials was too easy, while the use of the verbal instructions without the demonstrations was too difficult.

Though the seven revisions failed to result in an acceptable form of the test, it is still thought to be potentially useful because of its exceptional appeal to the children. It is also thought to have possibilities for other groups of retarded subjects. Limited experimentation suggested that presentation of the materials singly and in small groups would be suitable for younger severe retardates, while more complicated patterns seemed to be appropriate for the moderately retarded. The test was highly successful in holding the interest of both groups.

The *Bang Test* was also a test of imitation. The examiner struck the table an increasing number of times, saying, "Bang!" at each strike. The subject was then asked "to make bang!" or "to make bang! bang!" and so on. The same procedures were repeated for the transfer form, using the vertical surface of a wall instead of the horizontal tabletop. Various items in different orders were tried, the number of strikes ranging from one to seven.

The test presented several problems. A major difficulty was a marked tendency on the part of the subject to respond with random banging, a tendency which training did not often overcome. However, the most serious problem was that a number of the subjects disliked the test. Although they appeared to enjoy the demonstrations, they did not want to imitate them. In the case of subjects with whom the training was successful, another type of difficulty arose: some continued to strike or bang the material of subsequent tests, disrupting the testing situation.

The *Drum Test* and the *Tambourine Test* were attempts to retain the method of the preceding test. The items in the Bang Test were repeated, first by beating against a drum with the hand, and later by striking a tambourine. In both cases, the subjects did enjoy the test. However, their tendency to respond randomly was not lessened, and attempts at training tended to increase it still further.

The *Bell Test* was the most successful of the attempts to utilize the principles of the Bang Test. A domed white bell was used, in which a clapper hit the metal side of the bell when the projecting top was struck. The bell was very appealing to the subjects, though several revisions were necessary in its construction before it was suitable. A flat white disc raised well above the bell's dome was used in place of the original top, which was too small to attract the subjects' attention. More successful was a black disc, set high above the dome.

The use of the bell decreased the tendency to respond randomly, although it did not overcome it entirely. However, now that a greater number of correct responses were secured, another problem came into evidence. Subjects who had learned rudimentary counting performed better than subjects who had not. The test was therefore abandoned, since successful performance so evidently depended on differential past experience.

The *Egg Test* was a test of size discrimination. A lure was placed in the largest of a series of wooden eggs of various sizes. This egg was then presented to the subject in a different position for each item, the eggs being used in increasing numbers. Each egg was presented resting on its base, tapered end upward.

Since the first form of the test was found to be too easy, decreasing the differences between the sizes of the eggs was attempted. This change made the test too confusing, especially for the brain-injured subjects. Several attempts were also made to present the egg in a horizontal position. A gray board with oval indentations was used to prevent the eggs from rolling out of position. These attempts were not successful. When the eggs were presented so that the line from base to apex was parallel to the edge of the table, the test was too easy. On the other hand, the test became too difficult when either the tapered ends of the eggs or their bases were presented toward the subject. Several transfer forms were attempted, using black eggs of the same sizes as the white set, and various other shapes of corresponding relative sizes.

The test was finally abandoned as it became increasingly evident that the Buzz Test was a more suitable size discrimination test for this population, and was also easier to administer. However, the Egg Test was attempted in its first form with a group of younger severely retarded subjects, for whom the Buzz Test was too difficult. Results suggested that the Egg Test may be suitable for the younger subjects.

Vol. 74, No. 15 Whole No. 502, 1960

Psychological Monographs: General and Applied

SLEEP DEPRIVATION:
CHANGES IN PERFORMANCE AND PHYSIOLOGICAL INDICANTS OF ACTIVATION[1]

ROBERT B. MALMO AND WALTER W. SURWILLO[2]
Allan Memorial Institute, McGill University

EARLY investigations of sleep deprivation arose mainly from an interest in the influence of fatigue on performance. Except for a few studies such as the remarkable account of a 10-day vigil of a man who believed sleep to be unnecessary (Katz & Landis, 1935) this early work was rather inconclusive with respect to the effects of sleep loss on performance. A review of the early literature may be obtained from Edwards (1941), Katz and Landis (1935), Kleitman (1939), and Laslett (1928).

Not until recently (Kornetsky, Mirsky, Kessler, & Dorff, 1959; Wilkinson, 1958, 1959; Williams, Lubin, & Goodnow, 1959) have really adequate methodology and statistical treatment of results been applied to the problem of sleep loss and performance in human subjects (*S*s). Wilkinson's excellent studies are particularly noteworthy for the large number of *S*s employed and for the variety of the situations studied. One of his most significant findings appears to be the hitherto unrecognized importance of feeding back information to the sleep-deprived *S* about the scores that he is making. Without such feedback in a serial reaction test, performance was considerably impaired under the condition of sleep deprivation, whereas with feedback this impairment was practically eliminated. This finding is well established in two separate studies by Wilkinson (1958, 1959), and in the latter study, Wilkinson (1959) concludes that although inadequate incentive tends to yield poor performance even in normally rested persons, the need for adequate (and frequent) incentive becomes disproportionately great in those who have been sleep deprived.

Animal Studies. Employing rats as *S*s, the effects of enforced wakefulness on learning and performance in a water maze were investigated by Bunch, Frerichs, and Licklider (1938), and by Licklider and Bunch (1946). The 1938 experiments with periods of enforced wakefulness of 8, 16, 24, and 48 hours, produced rather surprising results, showing that in the case of all learning criteria (trials, time, and errors) the wakeful rats were superior to the controls. In the 1946 study, the animals' sleep was restricted to 4 hours per day, one group of wakeful rats undergoing 123 days of experimentation and a second group of wakeful rats undergoing 73 days of experimentation. Sleep deprivation again appeared to have facilitative effects although performance rather than learning appeared to be affected, and the authors stated (p. 346): "the better performance of the wakeful rats is largely to be accounted for in terms of their more aggressive behavior." In their conclusions the authors stressed that the restriction of the animals' sleep made them "highly irritable and highly reactive."

---

[1] Support for this research has come from the following sources: Medical Research and Development Division, Office of The Surgeon General, Department of the United States Army: Contract Number DA-49-007-MD-626; Defense Research Board, Department of National Defence, Canada: Grant Number 9425-04; National Institute of Mental Health, National Institutes of Health, United States Public Health Service: Grant Number M-1475; and National Research Council of Canada: Grant Number A. P. 29.

Grateful acknowledgment is made to A. Bartoshuk, J. Davis, L. Hart, L. Hutson, R. Quilter, D. Ross, M. Schnore, R. Stennett, and Elizabeth du Toit for assistance during the vigils. In addition, we wish to thank A. Bartoshuk for his constructive criticism of the manuscript.

[2] Present address: Gerontology Branch, National Institutes of Health, Baltimore, Maryland.

Nauta (1946), who produced sleeplessness in rats by making lesions in the preoptic nucleus, reported that after a period averaging 3 days "the exhausted animals fell into a state of coma which soon ended in death" (p. 303). It appears that sleeplessness produced by lesions may accelerate collapse, because Licklider and Bunch (1946) reported that rats in good health collapsed in 4 to 18 days without sleep. Before Nauta's rats collapsed, however, they too showed irritability, restlessness, and vigorous reactions to minor stimuli. Nauta likened some of this behavior to sham rage.

Sensory Thresholds. Several studies have been concerned with determinations of sensory thresholds under conditions of sleep deprivation. The length of the deprivation period varied from study to study, and in certain studies thresholds were determined after different periods without sleep. Cooperman, Mullin, and Kleitman (1934) in a study of cutaneous sensibility carried out with von Frey hairs, found that reports of pain occurred with weaker stimulation during a vigil of 60 hours. From their article it appears that this increased sensitivity to pain was not observed during the determinations made after 24 hours of sleeplessness; and this finding is in agreement with that of Schumacher, Goodell, Hardy, and Wolff (1940) that 24 hours' sleep deprivation produced no change in the thermal pain threshold. Except for the increased sensitivity to pain, and decreased sensitivity to sour taste (Furchtgott & Willingham, 1956), changes in sensory thresholds have not been observed following sleep loss. Negative findings were obtained with touch by Cooperman et al. (1934) after approximately 60 hours' deprivation, with audiometric tests by Goodhill and Tyler (1947) after 50 and 100 hours' deprivation, and with limens for salt and sweet taste by Furchtgott and Willingham (1956) after 24-, 48-, and 72-hour periods of sleep deprivation.

Electroencephalographic and Other Physiological Changes. EEG changes in the sleep-deprived *S* were, as far as we know, first studied by the group working with Kleitman. Among other things, they noted that when the *S* was engaged in a mental task requiring his concentration there was "a great discharge of beta waves which replaced the delta rhythm" (Blake, Gerard, & Kleitman, 1939, p. 55).

Tyler, Goodman, and Rothman (1947) investigated the effect of sleep deprivation on the percentage of time that each frequency occupied in the EEG tracing. These workers were careful to exclude from their measurements any segments of records taken when the *S* was suspected of dozing. The authors concluded that at 75 hours of sleep deprivation there was a shift in the percentage of time occupied by the various frequencies such that less time was occupied by frequencies around 9–10 cps, and more time was occupied by the higher frequencies (11–17 cps and 18 cps and higher). Control recordings were taken 3 days prior to the experimental period, and were compared with recordings taken under three separate sleep deprivation conditions: 50, 75, and 100 hours, with a different group of four *S*s serving in each of the three conditions. The authors concluded that the decrease in 9–10 cps activity and the corresponding increase in the higher-frequency activity was associated with the effort exerted by the *S*s in staying awake. This conclusion was based on their observation that similar changes in the EEG were found to occur while the *S*s were engaged in problem solving.

Bjerner's (1949) EEG study was mainly concerned with a different aspect of the sleep-deprivation phenomena. Instead of excluding EEG measurement during brief periods of sleep as Tyler et al. had done, Bjerner studied these inadvertent "catnaps." His five *S*s performed serial reactions with an apparatus specially designed for this study. The major part of the data analysis consisted in locating instances of gross lapse in performance (delayed actions) on the graphic tracings of performance, and then marking off the corresponding parts of the EEG and heart rate tracings for measurement. In this EEG analysis Bjerner concerned himself only with the number of alpha waves counted (*a*) during the period of delayed action, (*b*) for the 3 seconds before, and (*c*) the 3 seconds after the delayed action. Using the 3-second periods before and after delayed actions as basis for comparison, Bjerner found a significant drop in the number of alpha waves during delayed actions.

It is important to recognize that Bjerner's procedure was radically different from that of Tyler and his co-workers whose measurements were confined to parts of the vigil when the experimenters (*E*s) were certain that the *S* was not dozing. Bjerner's observations, on the other hand, were confined to periods when the *S* had dozed off during the performance of the serial reaction test. Bjerner's protocols actually noted the occurrence of snoring during some of the periods of delayed actions when the tracings to be measured were being taken. Accompanying the significant fall in number of alpha waves from the EEG during delayed actions was a significant transient fall in heart rate.

In a recent investigation Armington and Mitnick (1959) reported a progressive reduction in the amount of alpha (8–12 cps) activity with sleep deprivation. In the interpretation of these results, Armington and Mitnick referred to the work of Stennett (1957a) in pointing out that low alpha may be indicative of either a high or low level of activation, and they interpreted the reduction of alpha in their experiment as being due to drowsiness or a low level of activation.

Finally, an observation concerning the skeletal motor system may be mentioned. In a study in which he and his wife served as alternative *S* and *E*, Freeman (1932) noted increased quadriceps tonus on days following serious sleep loss.

Hallucinations. Numerous authors have mentioned hallucinatory experiences in sleep-deprived *S*s, but unfortunately the number of well described cases in which hallucinations definitely occurred is very small. With regard to hallucinations Tyler's (1947) investigation is unusually detailed. Of the total number of 588 *S*s studied, 444 *S*s were deprived over 60 hours (with 275 *S*s deprived as much as 112 hours). Some of the results of this study were complicated by administration of drugs (benzedrine sulphate and barbiturates) and by an extremely strenuous program of exercise. After stating that in almost all cases the psychological changes observed were only mild ones, Tyler wrote as follows:

> In a few cases the psychological changes which occurred were not so mild. These developed during the first 48 hours of sleeplessness and while they were still receiving placebos. The first such instance occurred just before dawn of the third day during a 112-hour experiment involving 35 men. This subject became extremely aggressive, developed delusions of grandeur (imagined himself on secret missions for the President, etc.), started unprovoked fights, became unmanageable, and had to be forcibly restrained. He was put to bed and slept through the entire day. On awakening he had no recollection of the events of the previous night. An examination of his record disclosed no indication of any psychotic tendencies during his training period; he was well liked, and considered quiet and reserved by his friends.
>
> Six other such cases occurred during the course of these studies: two were characterized by extreme irritability and aggressiveness requiring restraint; four cases by persistent auditory and visual hallucinations of such a nature that it was deemed desirable to have the men drop out. As in the case described above, none of these six men had a previous history of psychotic tendencies and, following sleep, all symptoms and all recollection of the preceding events disappeared (pp. 258-259).

As far as we know there has been no systematic study of physiological functions other than EEG and heart rate in conjunction with performance under conditions of sleep deprivation. Furthermore, with the possible exception of the work of Tyler et al. (1947), there have been no investigations made of physiological changes through the vigil under conditions which *force* the *S* to suppress sleep and to keep performance at or near the level of which he is capable in the absence of sleep loss.

The experiments to be described presently were designed to fill this gap. They contained feedback-of-information features to counteract lapses in the *S*'s tracking performance due to brief periods of sleep. Wilkinson's (1958, 1959) experiments showing the importance of feeding back information to a sleep-deprived *S* about his scores were noted earlier. Although the results of Wilkinson's experiments were not known during the design of this study, the *E* communicated with the *S* at the end of every trial and informed him what his score was on that trial and how that score compared with his best score up to that time. Two other means of informing the *S* concerning his errors (heaters on his legs that approached painful intensity with accumulation of gross errors, and the intensity of the tone in the earphones which was proportional to the *S*'s error) were employed to further decrease the liability of the *S* to lapses caused by a drift toward sleep.

To summarize, the main purpose of the present investigation was to observe changes in physiological and performance measures as a function of sleep deprivation under conditions designed to be highly alerting, viz., with the *S*s instructed to guard against falling asleep, even for brief intervals, and with various techniques involving continuous feedback of information concerning performance employed to keep alertness at a high level.

Data relevant to the matter of changes in pain thresholds with sleep deprivation were collected in the course of these experiments and finally, the relation of body sway to sleep deprivation was also studied.

METHOD

Subjects

Three young healthy male *S*s were employed in the experiments. S_1 was a 28-year-old clinical psychologist on our laboratory staff, S_2 a 26-year-old graduate student assigned to our laboratory for the summer, and S_3 a 25-year-old third-year medical student who had volunteered to serve as *S* after hearing about our experiments. Prior to his learning sessions on tracking before the vigil, this *S* had not been in our laboratory. S_1 and S_2 had served in pilot studies with very long sessions and they were highly practiced in tracking. All three *S*s were well paid for serving, and they all assumed their responsibilities with quite evident seriousness.

Medical examination prior to the experiment established that all were in excellent physical condition and, during the experiments, the *S*s were examined regularly by a physician. Except for a mild digestive upset in S_1 on the last day of his vigil (which, in the opinion of the physician did not require stopping the experiment) the experiments were uneventful from the medical point of view. S_1 was given one tablet of gelusil and one tablet (50 mg.) of gravol (dimenhydrinate) following breakfast after which his symptoms of mild nausea disappeared within an hour. At lunch time the same symptoms which were reported in the morning (mild nausea and lack of interest in food) returned. Approximately one-half of his lunch was left. These symptoms disappeared within one-half hour following administration of the same medication.

Physical examination revealed a hearing loss in S_3 which was subsequently investigated more closely because of its importance for auditory tracking. From comparisons of the *S*'s lower limen at 1000 cps with those of three observers with unimpaired hearing, it appeared that the *S*'s hearing loss for this frequency was in the neighborhood of 5 db. To compensate for this loss, the level of the tracking tone was raised approximately 5 db.

Apparatus

Tracking Apparatus

With certain modifications, the apparatus used in this study was identical with that previously described (Davis, Stennett, & Quilter, 1957). For this study, however, the directional cue was eliminated; hence an error in either direction produced a 1000-cycle tone in the earphones worn by the *S*. Seated comfortably in a semireclining position, the *S* was required to grasp and turn a knob with his right hand in such a manner as to prevent a 1000-cycle tone from appearing in the earphones. For purposes of EEG recording the *S* was blindfolded. The tracking task was a simple one in that it was only necessary for the *S* to turn the knob slowly to the left until he reached a mechanical stop and then let the knob return slowly to the right against the other stop. This pattern was repeated cyclically. As a further means of being sure that the *S* was constantly alert, he was asked to press his right foot down on a pedal each time that he came to one of the mechanical stops. By pressing down on the foot pedal the *S* closed a microswitch which caused a pip to appear on the polygraph tracing.

Time-off-target (TOT) was recorded by direct-reading SET clocks and also graphically by means of an integrator with a 4-sec. discharge rate. The instrument was calibrated such that 1 sec. off-target produced a deflection equal to 1 cm.; TOT was registered whenever the amount of error was approximately equal to or greater than 0.58°. Using the same kind of integrator, distance-off-target (DOT) in fractions of a degree was also recorded graphically. The amplitude of deflection per degree off-target was controlled by attenuators in the error-recording circuits. The attenuator settings were varied in accordance with the *S*'s level of performance. From calibrations, special scales were constructed for measurement of the DOT tracings.

Physiological Recording Apparatus

For recording physiological tracings and error data, three chart drives with ink-writing galvanometers were used. Chart drive no. 1 was part of a custom-built Edin EEG-EMG which recorded the primary tracings of EEGs and electromyograms (EMGs) (paper speed: 25 mm. per sec.). Electronic integrators, employing the condensor charge-discharge principle and connected in parallel with the galvanometers of this EEG-EMG unit, summated the muscle potentials over successive 4-sec. periods. The spike tracings from these integrators were recorded on chart drive no. 2 (paper speed: 1 mm. per sec.). High-pass filters designed to exclude frequencies below 20 cps were used for EMG integrations (where the modal frequency is of the order of 40 to 50 cps). For quantification of the EEG data two band-pass filters with fairly sharp cut-offs at the upper and lower limits of the 2–4 cps and 8–12 cps frequency bands were used in connection with integrators.

Chart drive no. 3 was used to record heart rate, respiration, systolic blood pressure, palmar conductance, DOT, and TOT. Heart rate was recorded by means of a modified Edin electrocardiotachograph (ECTG), and respiration was recorded with a Phipps and Bird pneumograph and recording tambour. An attempt was made to record systolic blood pressure continuously by means of an instrument copied with certain modifications from the design of Burns, Elliott, and James (1955). Unfortunately, however, the tracings from S_3 were not satisfactory and after Session 6, they were discontinued in the case of this *S*. Method of recording and measuring palmar conductance was similar to that described by Stennett (1957b).[3]

Bipolar sponge electrodes, soaked in electrode jelly and attached to the *S* by lastonet bands, were used for all EMG and EEG placements. Lead placements for S_1 and S_2 were as follows (EMGs from neck and biceps muscles were omitted for S_3): For the forehead-muscle recording, electrodes were placed 2 in. from the midline (4 in. from each other) and approximately 1 in. above the eyebrow. In recording from the muscles on the left side of the neck, the second cervical spine (found by counting from the prominent seventh cervical spine) was located and marked. A point 1⅛ in. outward from the second cervical spine (perpendicular to the backbone) was marked for the placement of the first electrode. The second electrode was placed 1¾ in. downward from the first electrode and ⅞ in. to the left of the midline.

For recording from the pronator teres muscle, a line was first constructed across the ventral surface of the arm from medial epicondyle to lateral epicondyle. The center of the first electrode was placed 1 in. medial to the midpoint between the two bony landmarks. The second electrode, its center 2 in. from the first, was the vertex of an equilateral triangle whose base lay along the elbow crease and whose vertex in the longitudinal plane lay along the midline of the arm (approximately 1¾ in. directly below the cubital fossa). In this study, recordings were taken from the left ("inactive") arm as well as from the right arm that was engaged in tracking.

In placing the first right-biceps lead, a point was marked one half of the distance from the anterior fold of the axilla to the cubital fossa, and the center of the first electrode was placed 1 in. proximally from this point (on the line between the two landmarks). The center of the second electrode was placed on the same line 2 in. distally from the center of the first electrode.

In placing the first electrode for the left leg, the distance from the head of the fibula to the lateral malleolus was measured, and the point which bisected this line was marked. The center of the first electrode was placed at this point. The center of the second electrode was placed on the same line, 2 in. proximal to the first.

Placements of all muscle leads except pronator teres are illustrated in a mimeographed manual by Davis (1952), and other technical points are dealt with in this manual. For EEG recording, bipolar sponge electrodes soaked in electrode jelly were attached to the *S* by lastonet tie bands, and these electrodes were placed two thirds of the distance from nasion to inion, and 3 cm. from the midline on each side. ECTGs were recorded by means of electrodes placed on contralateral limbs.

Heat-Pain Stimulation as a Means of Increasing the Subject's Motivation

In order to increase the *S*'s incentive to track well throughout the vigil, heater coils were connected to the legs which in turn were wired into a circuit activated by the error relay. Whenever the error score became disproportionately high, enough heat was generated by these coils to become actually painful. A detailed description of this apparatus has previously been published (Surwillo, 1958).

---

[3] Some important details of the apparatus for recording palmar conductance and some advances in recording technique are described in an unpublished paper by Malmo and Davis.

Other Apparatus

During part of the time that the *S* was not tracking, pain-threshold determinations were made and measurements of postural sway were recorded. The apparatuses employed in these two procedures were as follows:

Modified Hardy-Wolff Pain-Threshold Apparatus. Pain-threshold measurements were made with laboratory-constructed apparatus essentially the same as that used by Andrews (1943). The standard 3-sec. period of stimulation was provided by a shutter controlled in turn by a Hunter timer which was specially calibrated in our laboratory.

A 500-w., 120-v. Westinghouse T20 clear projection lamp (C13 construction) with medium, prefocused base, supplied the stimulus. This lamp was operated in the base-down position from the AC line through a voltage stabilizer and controlled by a Variac transformer. Each stimulus was measured by a wattmeter connected in the lamp circuit.

The lamp was mounted inside a bullet-shaped metal housing with an aperture of $\frac{13}{16}$-in. diameter in the nose. Two Bausch and Lomb lenses, 165.1 mm. in diameter, 31.9 mm. in thickness, and with 254.0-mm. (single lens) focus were combined (flat surfaces outward), and placed between the lamp and the aperture so that the focal length of the combined lenses was approximately 6 in. Two rounded plastic stops, 1 in. long and $\frac{1}{4}$ in. in diameter at their ends, protruded from the nose.

A very large area of the *S*'s forehead was blackened with Higgins drawing ink and allowed to dry well. After he had been informed concerning the threshold-taking procedure, the *S* placed his forehead against the stops in such a way that the focused light would fall on the blackened portion of his forehead. The *E* turned the wattage down to a low value and closed a switch which activated the Hunter timer. The timer caused the shutter to open and close allowing the light to fall on the *S*'s forehead for exactly 3 sec. For the first stimulation the wattage selected was one which almost always elicited the response "warm" from the *S*. Following his verbal response, the *S* withdrew his head from the stops while the *E* increased the Variac setting to a point where the wattmeter read 20 w. higher than on the previous stimulation. After 15 to 20 sec. the *S* was asked to place his forehead against the stops again, shifting the position slightly from where it had been on the previous stimulation. When the *S* was in position and had stated that he was ready, the *E* presented the second stimulus for 3 sec. and recorded the *S*'s response. This procedure was repeated with the *S* returning his forehead to the stops each time so that the area of skin stimulated was not the same on two successive occasions, and with the *E* advancing the intensity by 20 w. from stimulus to stimulus until the verbal response of pain was elicited from the *S*. It had previously been explained to the *S* that the procedure was by no means a test of his ability to endure pain, but rather was a means of determining what intensity of stimulation was required for him to perceive the first trace of pain. The *E* also explained that it was especially important for the *S* to concentrate during the latter part of stimulation because the slight pricking or stinging feeling, signifying that the pain threshold had been reached, appeared at the very end of the 3-sec. stimulation period. At least two threshold determinations were taken at each sitting but three determinations were invariably made unless the first two threshold values were extremely close.

All *S*s were given practice on the threshold-taking procedure and had thresholds taken prior to the vigil. During the vigil, thresholds were taken during the midafternoon of the first, second, and third days, and early in the morning (between 5:00 A.M. and 6:30 A.M.) of the second and third days. Pain thresholds were taken following sleep after the vigil in all cases. In consideration of the *S*'s convenience, however, it was possible to test only S_2 on the same morning after his sleep following the vigil.

Ataxiameter. This apparatus was an adaptation of Hull's (1933) Postural Movement Recorder. A clip with thread attached was fastened to the back of the *S*'s collar. This thread was led through a protective metal tubing on to a pulley and thence attached to an upright metal pointer whose base was held in a metal track. Next, the thread was led over a second pulley whence it ran vertically. A weight was attached to this end to keep the thread taut. When the *S* leaned forward, the pointer moved in its track to his right; when he leaned backward, the pointer moved to his left. The *S* was positioned initially so that the pointer rested midway along the track, pointing to zero on the scale. This scale, which extended 30 cm. to the right and 30 cm. to the left of the zero point, was marked off into major divisions of 1 cm. which in turn were quartered. In addition to the red pointer attached to the thread, two blue pointers were fitted into the track, one on each side of the red one. The blue pointers were used to record the maximum excursions of the red pointer (maximum forward and backward movement of the *S*).

The *S* stood with eyes closed and with feet in "V" position, at an angle of 30° with each other. After being instructed to keep his body as still as possible the *S* stood for 2 min., rested for 30 sec., and then stood for 2 min. again. The scale location of the red pointer was recorded every 10 sec. together with the maximum left and right (blue pointer) readings at the end of each 2-min. period.

TABLE 1

SCHEDULE OF TRACKING SESSIONS DURING VIGILS

Session No.	Time Scheduled	Day
1	11 A.M.	Tuesday
2	3 P.M.	
3	9 P.M.	
4	3 A.M.	Wednesday
5	11 A.M.	
6	3 P.M.	
7	9 P.M.	
8	3 A.M.	Thursday
9	11 A.M.	
10	3 P.M.	

Ataxiametric sessions were scheduled two hours apart throughout the vigils, and except for some occasions when delays occurred, these sessions were kept fairly close to schedule.

Procedures

From the *S*'s point of view the vigils were structured mainly by the 10 1-hour tracking sessions (see Table 1) and by meals. Five-min. tracking trials were employed, with 1-min. rest intervals between trials. There were 10 such trials in each experimental session. After each trial the *E* informed the *S* what his score was on that trial and how that score compared with his best score up to that time. In the case of S_2 and S_3, at the end of every session, full recordings were made during a 5-min. rest period in which the *S* remained seated in the chair under careful observation to make sure that he remained awake all through the period. In drawing up the master plan for all activities during the vigil, two hours prior to the times shown in Table 1 were allowed for precalibrations, the attachment of electrodes and other pickups to the *S*, and presession trial of all recordings. One hour was the minimum time for a session, but it was expected that most sessions would run slightly longer due to occasional stops to adjust or change faulty electrodes, or to make minor adjustments in the apparatus to ensure first-class recordings. A whole hour after the experimental session was set aside for removal of electrodes and attachments from the *S*, and for postcalibrations. Certain electrodes that could be worn for long periods without discomfort to the *S* were left on. Despite the large number of measures recorded it was possible to keep rather close to this schedule.

Meals from the institute's Diet Kitchen[4] were served to the *S* in the laboratory room. Breakfast was at 8:00 A.M., lunch at 12:30 P.M. and supper at 5:30 P.M. In addition to regular meals, morning coffee was served at 10:00 A.M. and light snacks were served at 2:00 P.M. and 10:00 P.M. The *S* remained in the same laboratory room throughout the vigil except for trips to the washroom and to a room adjacent to the laboratory to make an occasional telephone call. The experimental room was windowless and in order to prevent the *S* from perceiving the changes from daylight to darkness outside, he was blindfolded and led by one of the laboratory assistants whenever he went to the other two rooms. No attempt was made to conceal the time of day from the *S*; the purpose of the blindfold was merely to avoid giving the *S* opportunity to perceive changes in illumination from the sky. It has occasionally been reported that the subjective feeling of fatigue during a vigil is lessened at daybreak when the *S* perceives the sky becoming light.

Regularly following each tracking session two graphic rating scales were administered. In each instance the *S* was requested to place a vertical line somewhere along an unbroken horizontal line that was 20 cm. in length. The question guiding the rating of the first scale was, "How much effort did you have to put in the task in order to do as well as you could?" The guide phrases "considerable amount of extra effort," and "normal amount of effort" appeared at opposite ends of the horizontal line.

In the case of the second scale the question was, "How do you feel now?" Five guide phrases evenly spaced along the horizontal line were as follows: "very tired," "tired," "as usual," "fresh," and "very fresh." The order of the guide phrases from left to right was reversed on alternate presentations of each scale.

Soon after supper each evening and between midnight and 1:00 A.M. on Wednesday and Thursday, the *S* took a variety of tests administered by clinical psychologists. The purpose of these tests was mainly to keep the *S* actively engaged in order to prevent sleep. Between these planned activities and procedures the *S* was free to read, play games, or do whatever he pleased (S_1 spent much of his free time playing his guitar). The *S* was under constant observation by one member of the laboratory or another, and care was taken to have at least one well-rested laboratory assistant on duty at all times. At no time was the *S* left alone, even for a few minutes.

Treatment of Data

EEG

The EEG tracings recorded on one channel of the Edin instrument were used for monitoring purposes. Band-pass filters and integrators were employed to quantify the EEG tracings in the two frequency ranges of approximately 2–4 cps and 8–12 cps. The integrated data were separately recorded as deflections every 4 sec. (0.1 cm/sec paper speed).

---

[4] It is a pleasure to record our gratitude to the dieticians and to other personnel of the Diet Kitchen of the Allan Memorial Institute for their important contribution to this study.

Using all of the integrator deflections, the mean voltage per min. was determined by accumulating the number of millimeters in each successive series of 15 deflections dividing by the number of deflections (usually 15) and by reference to the specially calibrated ruler for the gain used in recording, reading off the corresponding microvolt value.

For each frequency range (2–4 cps and 8–12 cps) five mean 1-min. microvolt values were obtained from each tracking trial and 50 such values were obtained from one entire tracking session, providing no deflections were monitored out. The original tracing was carefully examined in order to ensure that no spike was measured if it covered a part of the tracing in which some artifact occurred. Where available (in the case of S_2 and S_3) the data for the 5-min. rest periods at the end of the sessions were treated in the same way as the data for the tracking trials.

Following measurement of the EEG integrations, the separate measurements were combined in various ways. The 250 1-min. mean microvolt values from the tracking trials in the first five tracking sessions were combined to yield a mean for the first half of the vigil, and similarly the mean microvolt value for tracking sessions in the second half of the vigil was also obtained. Mean microvolt values were also computed separately for each tracking session, for each 5-min. tracking trial, and for each 5-min. rest period.

EMG

With the exception of the tracings from the right arm, every sixth deflection was measured, commencing with the sixth deflection, such that 12 deflections in all were measured for every trial. For muscles of the right arm, in order to keep the task-induced tension approximately the same from measurement to measurement, the following deflections were measured: Numbers 4, 12, 19, 27, 34, 42, 49, 57, 64, and 72.

As in the case of EEG and other physiological measures, the main comparison was between mean values from the first half of the vigil with mean values from the second half.

Palmar Conductance

Measurements of palmar resistance were taken regularly every 15 sec. throughout each trial, commencing at the very beginning of the trial, so that there were 21 measurements per trial. Resistance values were immediately converted to conductance in micromhos with a table of reciprocals. All calculations and statistics were performed with conductance values as recommended by Woodworth and Schlosberg (1954, p. 140).

Respiration

Number of respirations counted to the nearest quarter respiration was obtained for the first 2½ min. of each trial, for the last 2½ min., and for the whole trial.

Heart Rate

Heart rates were determined from the ECTG tracing at 15-sec. intervals, these measures coinciding temporally with the measurements of palmar conductance.

Tracking Error Scores

TOT. The clock readings showed total TOT for each whole trial. Since the 4-sec. graphic integrations of TOT were available for fractionations involving parts of trials each deflection was measured with a rule calibrated directly in sec. Both TOT values were treated to yield a mean TOT score per trial for the first half of the vigil versus a mean TOT score per trial for the second half of the vigil. TOT scores were also obtained for the first half and last half of each trial.

DOT. Each 4-sec. integrator deflection was measured with a rule calibrated in degrees. Mean DOT scores for the first half and second half of the vigil were determined. DOT scores were also obtained for the first half and last half of each trial.

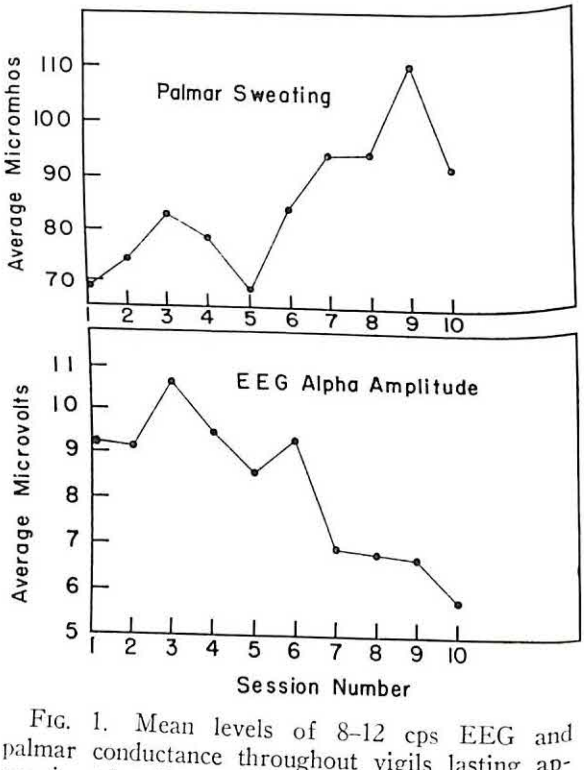


FIG. 1. Mean levels of 8–12 cps EEG and palmar conductance throughout vigils lasting approximately 60 hours. (Each point on curves represents mean value for 1-hour tracking session, combining the values for all three *S*s. Approximate time corresponding to each session number is given in Table 1.)

RESULTS

EEG, Palmar Conductance, and Respiration

Progressive EEG Changes during the Vigil in Relation to Concomitant Changes in Palmar Conductance and Respiration. The lower graph in Fig. 1 presents data for EEG alpha quantified by means of the 8–12 cps band-pass filter. The upper graph in Fig. 1 presents data from measurements of palmar conductance. Each point on the curves represents the mean value for one entire tracking session. Figure 2 presents the corresponding data from respiration rate. Table 2 presents a comparison between mean values for the first half of the vigil versus mean values for the second half in the case of each S, and the p values for the differences obtained.

Table 2 indicates that all three of these physiological measures showed reliable change for each S during the vigil, with palmar conductance and respiration higher during the last half of the vigil and amplitude of 8–12 cps EEG lower during the last half as compared with the first half of the vigil.

EEG data quantified by means of the 2–4 cps band-pass filter also yielded significant changes for all three Ss from first to last halves of vigil. S_1 and S_3 showed drops in amplitude from 5.3 to 3.9 ($p < .001$) and from 3.3 to 3.0 ($p < .01$) mean microvolts, respectively, while S_2 showed a rise from 3.7 to 4.0 mean microvolts ($p < .001$).

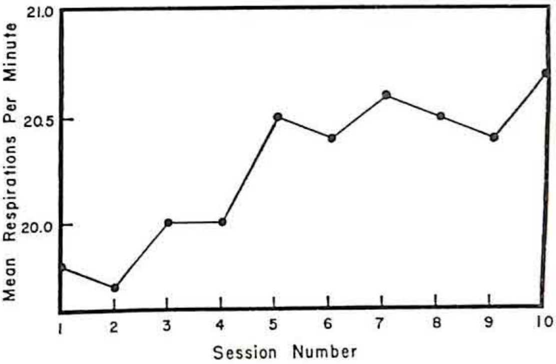


FIG. 2. Mean rates of respiration throughout the vigils. (Points on curves were obtained similarly to those in Fig. 1.)

Comparison of Within-Trial Changes: A Search for Lowered 8–12 cps EEG ("Alpha") Associated with Possible Drowsiness and Shift toward Sleep. Accompanied by a concomitant rise in palmar conductance and respiration, the finding of a progressive fall in 8–12 cps EEG activity is strong evidence that our Ss were very much awake and alert throughout the tracking sessions (Stennett, 1957a). However, to investigate this matter more closely, the data were subjected to further analysis.

An inspection of the 8–12 cps EEG data revealed that amplitude tended to fall during the trial, and this tendency seemed more marked in the second half of the vigil than in the first. We present the following analysis by Ss.

S_1: The mean difference between first and second halves of trials for the first five tracking sessions in the vigil was -0.16 μv.,

TABLE 2

CHANGES IN EEG 8-12 CPS AMPLITUDE, PALMAR CONDUCTANCE, AND RESPIRATION RATE DURING VIGIL

	S_1				S_2				S_3			
	1st Half of Vigil	2nd Half of Vigil	Diff.	p	1st Half of Vigil	2nd Half of Vigil	Diff.	p	1st Half of Vigil	2nd Half of Vigil	Diff.	p
8-12 EEG (μv.)	12.4	9.3	−3.1	$<.001$	6.8	5.1	−1.7	$<.001$	9.1	7.0	−2.1	$<.001$
PC (μ mhos.)	49.6	64.7	15.1	$<.001$	51.9	81.2	29.3	$<.001$	113.2	139.1	25.9	$<.001$
Resp. (per min.)	19.16	19.57	0.41	$<.05$	21.34	22.18	0.84	$<.001$	19.46	19.83	0.37	$<.02$

whereas the corresponding mean difference for the last five tracking sessions was 1.64 μv. The minus value for the first five sessions of the vigil indicated a mean rise in amplitude from the first to second halves of tracking trials. The difference between −0.16 and 1.64 μv. was a highly significant one ($p < .001$).

The data for palmar conductance and respiration were treated in a similar way as those for EEG. The difference values for palmar conductance were −1.08 and −0.28 for the first five sessions and the last five sessions, respectively. As is well known, palmar conductance characteristically falls during a trial of almost any kind of activity. The data indicated, however, that the fall was significantly less ($p < .05$) in the latter part of the vigil than it was in the first part of the vigil. Corresponding difference values for respiration rate were 0.51 and −0.29 for the first five and last five sessions, respectively. This difference was reliable at the .01 level, and indicated that in the last five sessions there was a shift in the direction of slowed respiration in the latter halves of the tracking trials. Inasmuch as this finding was the reverse of that from palmar conductance, the respiration tracings were studied. Careful examination revealed an irregularity in this *S*'s breathing pattern which seemed to be more prominent in the latter part of the vigil than in the first part. This irregularity consisted in the occurrence of deep breaths which had the effect of reducing the respiration rate. In the latter part of the vigil this *S* took more deep breaths than in the early part of the vigil. Thus in the first three sessions there were only 4 large respirations (two times the height of adjacent respirations and at least 2 cm. wide at the base) compared with 42 such deviating respirations observed in the last three sessions. The breakdown for first and second halves of these trials were 0 and 4, 11 and 31, for the two fractionations (Sessions 1–3 and 8–10), respectively. It would appear, therefore, that the larger number of these deviating respirations in the latter half of trials late in the vigil may have been responsible for the apparent reversal between the data from respiration and those from palmar conductance.

S_3: This *S* also showed greater within-trial fall of 8–12 cps EEG amplitude later in the vigil. The difference values were 0.82 and 1.76 and the difference between these two values was significant at the .05 level. However, neither palmar conductance nor respiration yielded significant differences when treated in a similar way.

S_2: Data from this *S* failed to show a significant EEG difference when fractionated in the way described. However, error scores (TOT and DOT) showed significant differences when treated in this way. For TOT the difference values were −3.61 and +1.87 for

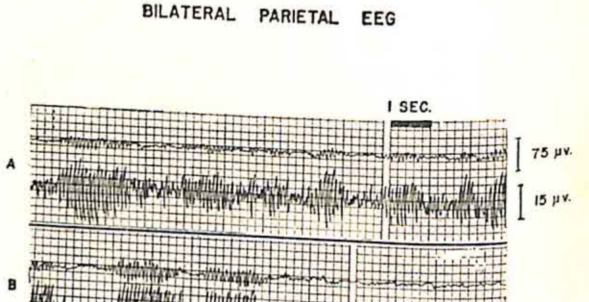


FIG. 3. Apparent disappearance of EEG alpha from the primary EEG tracing during brief episode of sleep in a young male *S* during tracking in an unpublished study by Malmo. (Search for similar phenomenon in the records of the present investigation failed. The arrow in this figure and those in Fig. 4 indicate the same moment in time.)

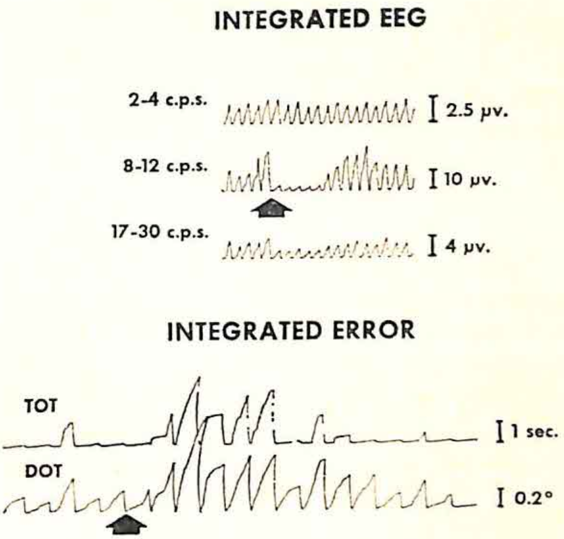
first and last parts of vigil, respectively (the plus score of 1.87 indicating a predominance of increasing error in the second half of trials), and the difference of 5.48 between these two values was highly significant (.001 level). For DOT the corresponding values were −0.009 and 0.003 and the difference of 0.012 between these two values was significant at the .05 level. Error scores from the other two *S*s were treated in this fashion but failed to yield reliable differences in either case.

Finally, it should be mentioned that a similar analysis was carried out on the heart rate data from the three *S*s but failed to yield any significant differences.

Tracings Illustrating Transition from Waking to Sleeping State. Finally, a careful search of the tracings was made for the purpose of locating instances of EEG changes characteristic of the kind which occur during a brief episode of sleep. This search did not reveal a single such instance.

In order to illustrate precisely what we were searching for, some tracings from an unpublished experiment by Malmo are reproduced in Fig. 3 and 4. These tracings were cut from Trial 9 of a series of tracking trials designed to bring the *S* to a practice level. The EEG lead placements were the same as those in the present study and the tracking apparatus was similar to the one described in the Methods section. As these tracings were being recorded, the *E* suspected from the pattern of changes occurring that the *S* had fallen asleep briefly during Trial 9. After the trial had been completed the *S* admitted upon questioning that he had indeed "dozed off."

Inspection of the record in Fig. 3 reveals that for a period of approximately 20 sec. there was virtual disappearance of the high-amplitude rhythmical waves with frequency close to 10 cps. This change is also seen in the integrator deflections in the upper part of Fig. 4 (second line). The deflections for the 17–30 cps component of the EEG (a band-pass filter for quantification of this component—Ross & Davis, 1958—was included in this study) also show a drop in amplitude (third line), while the 2–4 cps component appears to show no change. The integrator deflections for TOT (fourth line) and for DOT (fifth line) give evidence of a sudden increase in error that corresponds rather closely with the drop in amplitude of the EEG deflections. Because of the phase of the tracking cycle where the onset of sleep occurred, nothing should be made of the fact that the fall in EEG amplitude precedes the abrupt increase in error by several seconds. Heart rate (not included in the figure) showed a corresponding decrease during this brief period of sleep, followed by a rise.


INTEGRATED EEG

INTEGRATED ERROR

FIG. 4. Further data obtained from the same *S* as the one whose tracings are shown in Fig. 3. (The upper part of the figure shows the secondary EEG tracing consisting of the 4-sec. integrated outputs of three band-pass filters with the following frequency bands, from top to bottom: 2–4 cps, 8–12 cps, and 17–30 cps. The lower part of the record shows the sudden increase in error during the period when the *S* was dozing. Top line: 4-sec. integrations of TOT. Bottom line: 4-sec. integrations of DOT. The two arrows in this figure and the one in Fig. 3 indicate the same moment in time.)

Heart Rate

Mean heart rate for S_1 and S_3 was reliably higher (.001 level) in the second half of the vigil compared with the mean value for the first half of the vigil. The rises

TABLE 3

CHANGES IN MUSCLE TENSION DURING VIGIL

	S_1				S_2				S_3			
	Mean Microvolts				Mean Microvolts				Mean Microvolts			
Muscle	1st Half of Vigil	2nd Half of Vigil	Diff.[a]	p	1st Half of Vigil	2nd Half of Vigil	Diff.	p	1st Half of Vigil	2nd Half of Vigil	Diff.	p
Forehead	18.5	17.9	−0.6	*ns*	18.8	11.0	−7.8	<.001	9.3	10.8	1.5	<.01
Left neck	8.9	14.9	6.0	<.001	41.0	48.6	7.6	*ns*	—	—	—	—
Right biceps	42.6	29.9	−12.7	*ns*	26.3	33.8	7.5	<.01	—	—	—	—
L. pronator teres	11.0	5.9	−5.1	<.001	10.5	4.3	−6.2	<.001	12.4	8.7	−3.7	<.05
R. pronator teres	60.7	61.8	1.1	*ns*	76.9	41.9	−35.0	<.001	63.2	64.4	1.2	*ns*
Left leg	6.7	7.2	0.5	*ns*	6.6	5.9	−0.7	*ns*	2.9	3.7	0.8	*ns*

[a] Plus differences signify rise, minus differences signify fall in muscle tension from first to second half of vigil.

were from 90.5 beats per minute to 103.8 beats per minute for S_1 and from 81.6 to 86.7 for S_3. S_2 showed an opposite trend, a drop from a mean heart rate of 84.0 beats per minute for the first half of the vigil to 79.6 beats per minute for the second half of the vigil. The difference met the .05 criterion of significance.

EMG

Specificity of Tensional Increase during the Vigil. Table 3 presents the muscle tension data fractionated for first and second halves of vigil. A difference prefixed by a minus sign represents a fall in muscle tension from the first to the second half of the vigil. Although the interpretation of the EMG findings is hampered by the absence of data from neck and biceps muscles of S_3, inspection of Table 3 reveals an interesting pattern in the results. All *S*s showed a significant increase of tension in only one muscle. The particular muscle concerned varied from *S* to *S*: S_1 showed increased tension in the neck muscles, S_2 in the right biceps, and S_3 in the forehead muscle.

The neck muscle data from S_1 are plotted in Fig. 5. It will be seen that the steepness of EMG gradient as well as the level of muscle tension increased progressively throughout the vigil. As would be expected, the dotted line curve for the single (high-incentive) trial at the end of the vigil has a less regular appearance than the other curves; but it is of interest to note that although this curve has a steeper slope than the other curves, the level of muscle tension appearing at the end of this trial was approximately the same as the final value of tension in the curve based on mean values from the last three sessions.

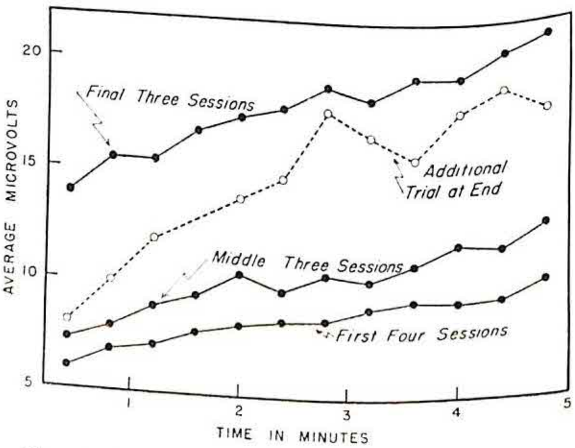


FIG. 5. Rise in neck muscle tension in S_1 during the vigil. (Data are plotted to permit visualization of mean EMG gradients for approximate thirds of the vigil and for final trial at the end of the vigil.)

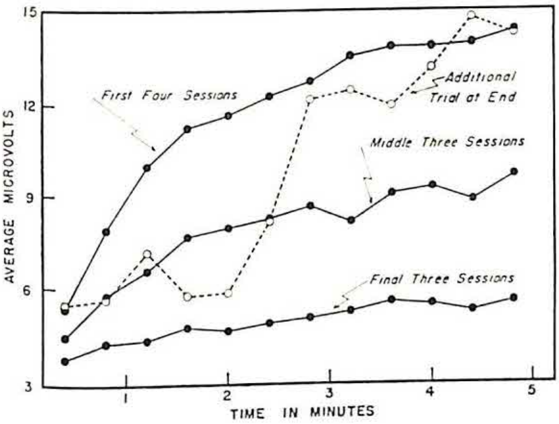


FIG. 6. Mean fall in left arm tension during the vigil. (Data are plotted in the same way as in Fig. 5, but data for all three *S*s are included.)

Progressive Fall in Left Pronator Teres Muscle Tension during the Vigil. Figure 6 presents curves for left pronator teres EMGs. These curves were plotted in the same manner as those shown in Fig. 5, but in this instance the mean muscle potential values plotted were taken from all three *S*s. As in Fig. 5 the curves show a progressive change in gradient steepness and in muscle tension level throughout the vigil, but the direction of change is exactly opposite to that shown in Fig. 5. There was continuous fall in level of tension and in gradient slopes for the left pronator teres muscle throughout the vigil. It should be noted that on the special incentive trial at the end of the vigil, mean tension and gradient steepness agreed very well with the corresponding mean values for the first four sessions. That the decrease of tension in this muscle was reliable in the case of each *S* may be seen by reference to Table 3.

Performance Data

Table 4 presents the data on tracking performance for each *S*. Amount of error for the first half of the vigil is compared with the amount of error for the last half. It is clear from Table 4 that the three *S*s were rather different with respect to the effects of the vigil upon their performances. S_1 showed clear impairment; the increases in TOT and DOT were both highly significant. On the other hand, S_2 showed an improvement in his TOT score that was also highly reliable. This *S*'s DOT score showed no change, however. Finally, S_3's change in TOT was not significant although his rise in mean error score in degrees met the .05 criterion of reliability. In summary, S_1 showed definite impairment, S_3 only slight impairment, while S_2 actually showed improvement in performance during the second half of the vigil compared with performance during the first half.

Subjective Reports Concerning Tracking Sessions

Hallucinatory Activity. At the end of the last (tenth) tracking session, in his own

TABLE 4

CHANGES IN TRACKING PERFORMANCE DURING THE VIGIL

	S_1				S_2				S_3			
	1st Half of Vigil	2nd Half of Vigil	Diff.[a]	*p*	1st Half of Vigil	2nd Half of Vigil	Diff.	*p*	1st Half of Vigil	2nd Half of Vigil	Diff.	*p*
Mean TOT (sec./trial)	64.8	120.8	56.0	<.001	98.2	86.9	−11.3	<.001	109.4	105.2	−4.2	*ns*
Mean DOT (degrees)	0.18	0.31	0.13	<.001	0.16	0.16	0.00	—	0.25	0.27	0.02	<.05

[a] Plus differences signify rise, minus differences signify fall in error scores (improvement) from first to second half of vigil.

handwriting S_1 wrote the following account of hallucinatory activity.

> [I noticed] a funny effect of [the] tone in [the] last session. Often I would completely forget the task, and the tones would become part of something like a dream. Different discrete tones (from switching the knob around) became people! And there were patterns of them: rows and squares, which were interesting but annoying, since I wanted to keep attention only on the tracking.

On the next day following a good night's sleep in the laboratory, the *S* added to his account as follows:

> I would have the feeling of moving, making my way among the "people" at this time, with no auditory sensations at all. I think during these times I was still moving the knob because Dave didn't say anything to wake me up.

As a psychologist who realized that under unusual conditions (e.g., perceptual isolation) hallucinations could appear in nonpsychotic individuals, S_1 was able to view these experiences rather objectively, and hence did not seem much disturbed by them.

S_2 did not have any visual hallucinations but he did report a loss of ability to appreciate the true shape of the knob by tactual and kinesthetic means. During one trial he just "blinked" at it to get a glimpse of the knob. He said that it was disturbing not to be getting sufficient information from kinesthetic and cutaneous senses. This loss of cutaneous appreciation of the knob apparently occurred mainly during Session 5 when the rotational action of the knob had become somewhat stiff. It seemed to the *S* that during this time he exerted considerably more pressure on the knob than at other times.

S_3 reported hallucinations which were inanimate, and were a materialization of a visual schema that he had been employing as an aid in tracking. He visualized a rectangle which represented the error-free zone in tracking. Inside this rectangle he visualized a black dot which represented the position of the brush that he could move by means of the knob that he held in his right hand. In his mind's eye he visualized the rectangle as moving from right to left as tracking commenced, and his self-instructions were to turn the knob in such a way as to keep the dot within the rectangle.

In the practice trials when he was learning to track, the mechanism of the tracking apparatus was explained to him by means of a diagram, and being "visual minded" (the *S* customarily used visual aids in studying for examinations), he projected the visual image as if seen on a piece of paper. Although these images appeared to have clarity verging on the eidetic they differed in certain important respects from the hallucinatory images that occurred in the later sessions. Unlike the hallucinations, the visual images used as guides had definite form and they were completely under control.

The transition from the stage of imagery used as visual aid to hallucinations was a smooth one. Apparently the hallucinations first appeared briefly in Session 8 but did not develop fully until Session 9. The *S* began to lose control of his imagery and the definite form of the images was replaced by undulating, wavy, irregular borders without sharp angles. The hallucinated figures were two-dimensional ones, projected some distance from the eyes as if on a movie screen. In his notes the *S* wrote as follows:

> The test [Session 9] began in much the usual way, but during trial no. 3 or no. 4 the visual image I have of the tone and which I use in tracking was replaced by indistinct hallucinations—halves of large objects—and the tracking test became a test of outlining these objects. I was rather surprised to note that my score on the first such test, trial no. 3 I think, was as good as it was, or as [the *E*] said it was. Deep respiration seemed to overcome or help me to dispel these mental imageries, and it (deep breathing) had to be maintained voluntarily and with considerable concentration in order to permit me to visualize the test pattern as I have been used to.

Further questioning revealed that for a session or two the *S* was preoccupied with what he thought was a connection between the hallucinations and breathing. This even developed for a brief time into a delusion that the hallucinations were being induced by the *E*s. He thought that he noted some connection between the disappearance of the hallucinations and "disturbances" in the control room where the *E*s were operating the equipment. By "disturbances" the *S* meant talking. The *S* had the feeling that his deep breathing was undesirable from the point of view of the *Es* (because of the artifacts produced in the physiological recordings), but he stated in his notes that he continued to breathe deeply despite this. Careful examination of the respiratory tracings revealed very little actual deep breathing during tracking trials, although increased amplitudes of the respiratory tracings were noted in the rest periods between trials in Session 9, and later in this session the pattern of respiration which had previously been very regular became irregular. It is likely, therefore, that the *S* exaggerated the amount of deep breathing that he engaged in.

Again from the *S*'s notes there were the following statements made immediately following Session 9 when the hallucinations were at their peak:

> During the tracking there were gross tremors which later became smaller to fine, and which are only now disappearing; also there was excess salivation.

After the final session (10) the *S* wrote as follows:

> I put maximum effort into concentrating on sounds alone, in an effort to reduce the number of hallucinations, and I very seldom allowed myself consciously to project a visual image of the sound. I saw only a few hallucinations which were similar to those described above; when they occurred they were transitory and my surprise or fear of seeing them was changed as a result of the [explanation of how hallucinations may be produced in normal individuals under unusual circumstances].

The *S* emphasized that he seemed to be able to ignore the hallucinations in this session by concentrating on the sound only.

It is interesting to note that roughly six weeks later when this *S* was interviewed again there appeared to be some confusion in his mind about the details of the visual schema that he had used during some nine hours of tracking during the vigil.

Rating Scales. Figure 7 presents the data from the rating scales for effort ("How much effort did you have to put in the task in order to do as well as you could?") and feeling of fatigue ("How do you feel now: very fresh, fresh, as usual, tired, very tired?"). The curves, which represent mean values for the three *S*s, show the expected changes in the direction of increased effort, and presence of a tired feeling in the latter part of the vigil. S_3 appeared to show a "second wind" phenomenon during the sixth session which is reflected in the group curve. Examination of the physiological and performance data did not confirm this finding. Except for a steeper EMG gradient for the left pronator teres muscle during this session, there were no conspicuous differences. The gradient in the right pronator teres was quite low, and there was nothing to indicate a break in trend in the data for heart rate, respiration, palmar conductance, and the error scores (DOT and TOT).

Further Subjective Reports from Protocols. S_1 found it more difficult to stay awake during the tracking sessions than at any other time. Early in the vigil he stated that he was bored and found it more difficult to overcome boredom than to stay awake. As the vigil continued to the final tracking ses-

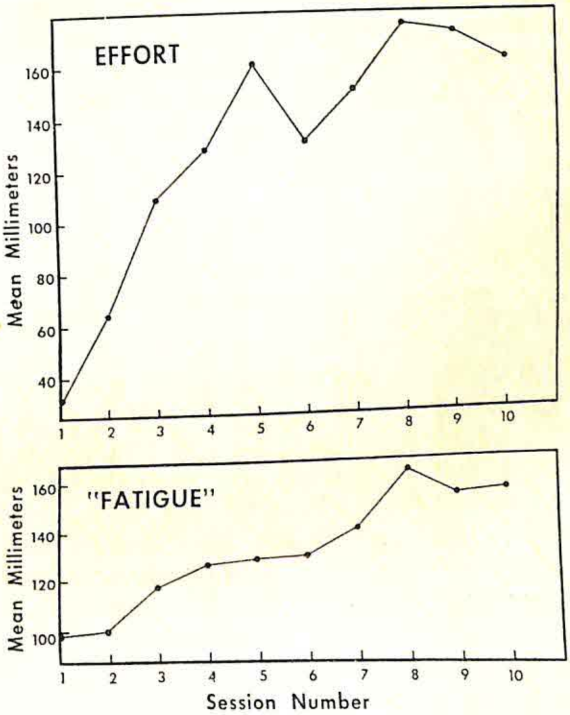


FIG. 7. Mean rise in ratings of effort and feeling of fatigue.

sion, however, his notes indicate that it was increasingly difficult to stay awake and that he was constantly fighting sleep. Verbatim notes from Thursday at 4:15 A.M.:

> Very hard to stay awake during tracking and almost impossible to keep eyes shut in between trials: It's either open them or sleep.

This *S* had the impression that by the eighth tracking session he was "getting lost" in the second half of the trials. Analysis of the error data, however, failed to support this subjective impression.

From his notes, it appeared that S_2 was more conscious of drowsiness when he was not occupied (as during lead placement and rest periods). When tracking commenced his feelings of drowsiness were reduced. S_2's notes reflect a strong desire for high achievement in this task. He wrote that his performance had "deteriorated" during Session 5, but analysis of the error scores revealed that mean error score for this session was actually lower than the scores had been in the previous four sessions. Following Session 8 he wrote:

> It seemed that even though my thoughts were elsewhere, I was able to track automatically to a greater or lesser degree, being only minimally conscious of the tone.

In the last sessions S_2 reported "catnaps" and discouragement regarding his performance, but analysis of the error data revealed that he overestimated the effects of sleep loss on his performance.

Except for the reports on hallucinations, there was little else of interest in the notes from S_3.

Pain Thresholds

In each case the mean pain threshold for the last day of the vigil was lower than the mean pain threshold for determinations made prior to the last day, and following sleep after the vigil. The mean thresholds in watts were 360 and 401 for S_1 (difference of 41 significant at the .02 level), 280 and 290 for S_2 (difference of 10 not significant), and 245 and 277 for S_3 (difference of 32 significant at the .001 level).

Ataxiameter Data

The ataxiameter data in Table 5 show that body sway was significantly increased in the second half of the vigil in the case of S_1, significantly decreased in the case of S_2, and not significantly changed in the case of S_3. The decrease in sway for S_2 was significant for one measure (maximal shift) at the .02 level, and the decrease by the other measure (mean shift per 10 sec.) barely missed significance at the .05 level.

TABLE 5

CHANGES IN BODY SWAY AS MEASURED BY THE ATAXIAMETER DURING THE VIGIL

	S_1				S_2				S_3			
	1st Half of Vigil	2nd Half of Vigil	Diff.[a]	p	1st Half of Vigil	2nd Half of Vigil	Diff.	p	1st Half of Vigil	2nd Half of Vigil	Diff.	p
Mean shift/10 sec. (cm.)	0.57	1.30	0.73	$<.02$	0.31	0.23	−0.08	*ns**	0.22	0.26	0.04	*ns*
Mean maximal shift/2 min. (cm.)	4.93	8.91	3.98	$<.05$	3.30	2.04	−1.26	$<.02$	2.48	2.26	−0.22	*ns*

[a] Plus differences signify increase, minus differences signify decrease in amount of body sway from first to second half of vigil.

* $t = 2.06$, $t_{.05} = 2.07$.

DISCUSSION

EEG

The decrease in amplitude of the alpha component of the EEG observed with progressive sleep loss in this study is consistent with the findings of previous investigators. The methods used in quantifying the EEG data have, of course, differed from investigator to investigator. Tyler et al. (1947) used manual methods of analysis and found a decrease in the 9–10 cps band. Bjerner (1949) also used manual methods but classified frequencies of approximately 10 cps as alpha activity in the case of three of his *S*s and 8.5 cps and 7.5 cps, respectively, in the case of the other two (pp. 58–59). Armington and Mitnick (1959), on the other hand, used an automatic digital-type method with time as the unit of measurement. In this case, the frequency band 8–12 cps was quantified and EEG activity was considered to be present only when the amplitude of the activity in this band exceeded 25 μv. (with 2 *S*s, 15 μv. was the criterion). Despite these obvious differences, however, there has been (together with the present results) substantial agreement in the finding that EEG activity in the "alpha" band has decreased with prolonged deprivation of sleep.

Such agreement does not, however, extend to the interpretation of these findings. Tyler et al. concluded that their results were associated with the effort of the *S*s to stay awake. Bjerner, on the other hand, concluded quite reasonably that his own findings of decreased alpha activity during sleep loss were associated with brief periods of sleep. Finally, Armington and Mitnick interpret their findings of low alpha during sleep deprivation as being due to drowsiness.

In previous articles (Malmo, 1958, 1959), some of the data reported in the present experiment were dealt with in relation to activation theory. To summarize the main points previously made, depriving an organism of sleep appears to be like depriving it of other basic needs. Under conditions of sufficient environmental stimulation this deprivation leads to increased activation. That is, judging from the overall picture of physiological change during our experimental vigils, it was concluded that the tracking task was sufficiently stimulating to interact positively with deprivation to produce a significant rise in activation during the 60-hour vigil.

It is clear that a difficult problem of interpretation has been raised by these different studies. Before returning to a consideration of the separate studies, it will be useful to recognize one very important general point. This is simply that low alpha EEG activity can, when it is observed, indicate either a high level of alertness (high activation) or a drift toward sleep (low activation). This to be sure is not a new finding—it has been reported in numerous studies. Only recently, however, have the implications of this finding for behavioral studies been made explicit (Malmo, 1959; Stennett, 1957a). We may conclude then that, if we consider only the level of alpha EEG activity at a given time—and nothing else—the question of whether an *S* shows high or low activation is indeterminate. In the light of this conclusion it is obvious that the apparent issue of interpretation which has been raised reduces to a matter of other evidence. It revolves around the question: Besides level of EEG alpha, what *other* evidence has been brought to bear on the matter? Let us examine the various studies in turn with reference to such additional evidence.

Although no measures other than the EEG were employed by Tyler et al. (1947) to assist with the interpretation of the alpha findings, these workers did show that accompanying the fall in 9–10 cps activity there was a rise in the fast-frequency component of the EEG (i.e., a higher percentage of waves in the 11–17 cps band, and in the band from 18 cps and higher). This certainly is more consistent with the interpretation of increasing activation rather than the reverse inasmuch as a decrease *not* an *increase* in the higher frequencies is to be expected with a change in the direction of sleep (Lindsley, 1957, p. 68).

Bjerner (1949), it will be recalled, did not deal with long-term changes in alpha but with short-term, transient changes during "delayed actions." His finding of decreased alpha was associated with evidence from behavioral observations as well as evidence from another physiological measure (heart rate). The EEG changes he observed were accompanied by overt signs of sleep (marked delay in response to a signal, snoring) along with a reliable fall in heart rate. From this additional evidence it seems clear that Bjerner's *S*s were at these times drifting down the activation dimension.

It is unfortunate that Armington and Mitnick (1959) presented no physiological evidence besides the 8–12 cps EEG data; but it seems likely that they are correct in interpreting their findings of low alpha during sleep deprivation as being due to drowsiness. In other words, it appears probable that they were unsuccessful in their attempt to investigate the effects of sleep deprivation under conditions less favorable for dozing than in previous studies: having their *S*s stand was certainly no guarantee that the *S*s would not drift toward sleep.

To reiterate, in our experiment the EEG and other physiological findings are consistent with the interpretation that our *S*s were alert and in a high state of activation. We are not aware of any instance where rise in palmar conductance or respiration rate has been reported to coincide with a drift down the activation dimension toward sleep.

Because so little has been reported about the 2–4 cps component of the EEG in waking *S*s it is difficult to interpret the finding that tracings from S_1 and S_3 showed a significant fall in amplitude. However, two rather general comments may be made. Firstly, it is interesting to note that once again it was S_2 who was the exception. Secondly, S_1 whose tracking performance deteriorated more than the others showed the greatest fall in 2–4 cps EEG. If S_1's drop in performance were interpreted as due to drift toward a sleepy state, it is likely that, if anything, a rise in amplitude of frequencies in this low range would have been expected instead of the fall that was observed (Malmo, 1959, p. 380). The same general argument holds for S_3 who showed a slight drop in performance. S_2, it will be recalled, actually improved in tracking in the second half of the vigil so that the slight rise in 2–4 cps EEG amplitude can hardly be accounted for in terms of increasing drowsiness in this *S*.

Other Physiological Measures

Heart Rate. S_1 and S_3 showed significant rise in heart rate during the vigil. This finding was consistent with the rise in palmar conductance and in respiration rate. The results from S_2, on the other hand, were the exact opposite; in his case there was falling heart rate which on the basis of a comparison between first and last halves of the vigil was statistically significant.

This reversal in heart rate is difficult to interpret considering that the palmar conductance, respiration rate, and 8–12 cps EEG data were very similar to those of the other two *S*s (see Table 2). However, we may note that, in referring to the experiments of Hickam and his co-workers, Gellhorn (1953, p. 338) indicated that under activating conditions which produce heart rate acceleration in the majority of *S*s, some *S*s show diminished heart rate and lowered blood pressure. As a matter of fact, S_2 did show a fall in blood pressure from the first half to the second half of the vigil, and although the difference was not large it was significant (at the .01 level). S_1, on the other hand, showed a rise in blood pressure from the first half to the second half of the vigil, though this rise was not statistically significant. The attempt to record blood pressure from S_3 was abandoned because finger movement artifacts made it impossible to use this instrument (which depended on a finger pickup).

However, it should be made clear that there is evidence against the notion that S_2 simply reacts with a drop in heart rate under all conditions in which the heart rates of others increase. Data taken several months before his participation in the sleep deprivation experiments clearly showed typical heart rate acceleration in this *S* under certain activating conditions. Comparison of heart rate for 5-min. trials during the last half of the session with that for the 5-min. trials during the first half of the session showed a significant rise in the second half.

Finally, it may be noted that even in the second half of the vigil despite the downward shift in the heart rate baseline, heart rate rose regularly at the end of each trial, during the time that the *S* was waiting for and receiving information concerning his performance on that trial.

EMG. As previously stated each *S* showed a significantly *increased* tension in one particular muscle (left neck, right biceps, and forehead, for S_1, S_2, and S_3, respectively), and all *S*s showed a progressive *decrease* in level and in steepness of gradient of EMGs from the left pronator teres muscle during the vigil (see Fig. 6).

The findings of increased muscle tension localized in one muscle group has been discussed previously in relation to the specificity of localized muscle tension in psychiatric patients with complaint of discomfort from these tense muscles (Malmo, 1959). The previous statement was as follows:

> Where high level activation is long continued as in a vigil or in certain psychoneurotic patients, it appears that skeletal tension may become localized to a single muscle group. The discomfort associated with this tension in some patients can become extremely severe. It should be noted that in one-session experiments, where rise in activation was for relatively short intervals of time, tensional rise occurred in more than one muscle group.
>
> Methodologically, these results are important because they reveal a difference between EMGs and some other physiological measures with respect to gauging activation. Unlike heart rate or respiration rate that invariably yields one measure no matter how it is recorded, there are as many measures of muscle tension as there are muscles that can be recorded from. It appears that when sufficient care is taken, EMGs may be very valuable in helping to gauge activation, but that considerable caution is required in the interpretation of results, and especially in the interpretation of negative results (p. 383).

At the very end of the vigil each *S* was given an extra 5-min. trial, after being informed that this was his *very* last tracking trial. Under these conditions it was interesting to note that the tension in the left pronator teres muscle that had been gradually decreasing all through the vigil, returned to the level it had reached during the early trials and also showed approximately the same steepness of gradient. This finding suggests that there might have been something like a conservation of energy mechanism operating to reduce the tension in the left arm during the vigil, and that when the vigil was over at last, and such conservation of energy was no longer required, the tension in the left arm returned to its previously higher level. A puzzling question is why, if such a conservation mechanism existed, it did not operate on all of the muscles. Of course with a larger group of *S*s and with more channels of EMG it is quite possible that variations would be found, with some individuals showing rise in more than one muscle, and some individuals perhaps not showing a rise in any muscle. However, considering the small number of *S*s in this investigation, the consistency of the findings was remarkable.

Pain Thresholds

The three *S*s were consistent in the direction of change (a drop) in pain threshold with sleep deprivation, although S_2's change was not statistically significant. These findings are on the whole consistent with those of Cooperman et al. (1934) whose report was summarized in the introductory section of this paper. Although the negative results of Schumacher et al. (1940) are occasionally cited as evidence against a change in pain threshold with sleep deprivation, this does not appear to be a fair conclusion because the threshold determinations were carried out after only 24 hours of sleep deprivation. Actually the main purpose of the experiments by Schumacher and his coworkers was to test the influence of a rather moderate amount of sleep deprivation on the pain threshold.

As far as we know, pain is the only modality that has been found more sensitive under conditions of sleep deprivation. In the introductory section we briefly summarized the findings concerning other sense modalities. From this evidence it appears that pain is unique with regard to the effect sleep deprivation has upon it. This uniqueness is of interest in relation to Hebb's (1949) theory of pain. Hebb emphasizes

SUMMARY AND CONCLUSIONS

Three separate 60-hour vigils were held with each of three healthy young male *S*s. During each vigil, 10 1-hour recording sessions were held several hours apart. In each session the following measurements were recorded continuously while the *S* was performing on a tracking task with 5-min. trials and 1-min. rests: EEG, EMGs from various muscles, respiration, palmar conductance, and heart rate.

Main findings were that in all three *S*s, palmar conductance and respiration showed a statistically reliable increase, and 8–12 cps EEG a reliable decrease in amplitude. Heart rate was reliably increased in two of the three *S*s, (though reliably decreased in the third *S*). In addition, each *S* showed reliably increased muscle tension in only one muscle group (a different one in each *S*).

On the whole the data strongly supported the conclusion that sleep deprivation had the effect of increasing level of activation. In this respect the present investigation appeared to be different from certain other investigations of sleep deprivation in which conditions were not sufficiently alerting to prevent the *S*s from drifting toward sleep. With the overall indication of increasing activation from the peripheral measures, the fall in 8–12 cps EEG was interpreted as also indicative of increasing activation, and not as drift toward sleep.

Under conditions of this investigation, increased muscle tension appeared to be localized in one muscle group rather than generalized over the entire body. In fact, tension in some parts of the body actually fell progressively during the vigil.

It appeared that the special conditions of the tracking sessions were particularly conducive to the appearance of hallucinatory experiences.

REFERENCES

Armington, J. C., & Mitnick, L. L. Electroencephalogram and sleep deprivation. *J. appl. Physiol.*, 1959, **14**, 247–250.

Bexton, W. H., Heron, W., & Scott, T. H. Effects of decreased variation in the sensory environment. *Canad. J. Psychol.*, 1954, **8**, 70–76.

Bjerner, B. Alpha depression and lowered pulse rate during delayed actions in a serial reaction test. *Acta physiol. Scand.*, 1949, **19**, Suppl. 65.

Blake, Helen, Gerard, R. W., & Kleitman, N. Factors influencing brain potentials during sleep. *J. Neurophysiol.*, 1939, **2**, 48–60.

Bliss, E. L., Clark, L. D., & West, C. D. Studies of sleep deprivation: Relationship to schizophrenia. *AMA Arch. Neurol. Psychiat.*, 1959, **81**, 348–359.

Bunch, M. E., Frerichs, J. B., & Licklider, J. R. An experimental study of maze learning ability after varying periods of wakefulness. *J. comp. Psychol.*, 1938, **26**, 499–514.

Burns, B. D., Elliott, H., & James, C. E. Continuous recording of respiration, heart rate, and systolic blood pressure. *Canad. Serv. med. J.*, 1955, **11**, 509–514.

Cooperman, N. R., Mullin, F. J., & Kleitman, N. Studies on the physiology of sleep: XI. Further observations on the effects of prolonged sleeplessness. *Amer. J. Physiol.*, 1934, **107**, 589–593.

Davis, J. F. *Manual of surface electromyography.* Montreal: Laboratory for Psychological Studies, Allan Memorial Institute, McGill University, 1952. (Mimeo.)

Davis, J. F., Stennett, R. G., & Quilter, R. E. An auditory tracking device designed for use in conjunction with continuous EEG recording. *Percept. mot. Skills*, 1957, **7**, 239–244.

Edwards, A. S. Effects of the loss of one hundred hours of sleep. *Amer. J. Psychol.*, 1941, **54**, 80–91.

Freeman, G. L. Compensatory reinforcements of muscular tension subsequent to sleep loss. *J. exp. Psychol.*, 1932, **15**, 267–283.

Furchtgott, E., & Willingham, W. W. The effect of sleep-deprivation upon the thresholds of taste. *Amer. J. Psychol.*, 1956, **69**, 111-112.

Gellhorn, E. *Physiological foundations of neurology and psychiatry.* Minneapolis: Univer. Minnesota Press, 1953.

Goodhill, V., & Tyler, D. B. Experimental insomnia and auditory acuity. *Arch. Otolaryngol.*, 1947, **46**, 221–224.

Hebb, D. O. *The organization of behavior.* New York: Wiley, 1949.

Hebb, D. O. The motivating effects of exteroceptive stimulation. *Amer. Psychologist*, 1958, **13**, 109–113.

Heron, W. The pathology of boredom. *Scient. Amer.*, 1957, **196**, 52–56.

Heron, W., Doane, B. K., & Scott, T. H. Visual disturbances after prolonged perceptual isolation. *Canad. J. Psychol.*, 1956, **10**, 13–18.

Hull, C. L. *Hypnosis and suggestibility.* New York: Appleton-Century, 1933.

Katz, S. E., & Landis, C. Psychologic and physiologic phenomena during a prolonged vigil. *Arch. Neurol. Psychiat.*, 1935, **34**, 307–316.

Kleitman, N. *Sleep and wakefulness.* Chicago: Univer. Chicago Press, 1939.

Kornetsky, C., Mirsky, A. F., Kessler, Edith K., & Dorff, J. E. The effects of dextro-amphetamine on behavioral deficits produced by sleep loss in humans. *J. Pharmacol. exp. Ther.*, 1959, **127**, 46–50.

Laslett, H. R. Experiments on the effects of the loss of sleep. *J. exp. Psychol.*, 1928, **11**, 370–396.

Licklider, J. C. R., & Bunch, M. E. Effects of enforced wakefulness upon the growth and the maze-learning performance of white rats. *J. comp. Psychol.*, 1946, **39**, 339–350.

Lindsley, D. B. Psychophysiology and motivation. In M. R. Jones (Ed.), *Nebraska symposium on motivation, 1957.* Lincoln: Univer. Nebraska Press, 1957. Pp. 44–105.

Malmo, R. B. Measurement of drive: An unsolved problem in psychology. In M. R. Jones (Ed.), *Nebraska symposium on motivation, 1958.* Lincoln: Univer. Nebraska Press, 1958. Pp. 229–265.

Malmo, R. B. Activation: A neuropsychological dimension. *Psychol. Rev.*, 1959, **66**, 367–386.

Morris, G. O., Williams, H. L., & Lubin, A. Misperception and disorientation during sleep deprivation. *AMA Arch. Neurol. Psychiat.*, in press.

Nauta, W. J. H. Hypothalamic regulation of sleep in rats: An experimental study. *J. Neurophysiol.*, 1946, **9**, 285–316.

Ross, W. R. D., & Davis, J. F. Stable band-pass filters for electroencephalography. *IRE Canad. convention Rec.*, 1958, No. 860, 202–206.

Schumacher, G. A., Goodell, Helen, Hardy, J. D., & Wolff, H. G. Uniformity of the pain threshold in man. *Science*, 1940, **92**, 110–112.

STENNETT, R. G. The relationship of alpha amplitude to the level of palmar conductance. *EEG clin. Neurophysiol.*, 1957, **9**, 131–138. (a)

STENNETT, R. G. The relationship of performance level to level of arousal. *J. exp. Psychol.*, 1957, **54**, 54–61. (b)

SURWILLO, W. W. A new method of motivating human behavior in laboratory investigations. *Amer. J. Psychol.*, 1958, **71**, 432–436.

TYLER, D. B. The effect of amphetamine sulfate and some barbiturates on the fatigue produced by prolonged wakefulness. *Amer. J. Physiol.*, 1947, **150**, 253–262.

TYLER, D. B., GOODMAN, J., & ROTHMAN, T. The effect of experimental insomnia on the rate of potential changes in the brain. *Amer. J. Physiol.*, 1947, **149**, 185–193.

VERNON, J. A. Pain thresholds and time orientation in sensory deprivation. Paper read at Symposium "Experimental Foundations of Clinical Psychology," University of Virginia, April 1959.

WILKINSON, R. T. The effects of sleep loss on performance. *Cambridge Appl. Psychol. Res. Unit Rep.*, 1958, No. 323/58.

WILKINSON, R. T. Repeated measurements of the effect of lack of sleep on the same subjects. Report prepared for the Operational Efficiency Subcommittee of the RNPRC, September 1959.

WILLIAMS, H. L., LUBIN, A., & GOODNOW, JACQUELINE J. Impaired performance with acute sleep loss. *Psychol. Monogr.*, 1959, **73**(14, Whole No. 484).

WOODWORTH, R. S., & SCHLOSBERG, H. *Experimental pschology.* New York: Holt, 1954.

(Received March 18, 1960)

Vol. 74, No. 16 Whole No. 503, 1960

Psychological Monographs: General and Applied

ABILITY FACTORS AND COMPONENT PERFORMANCE MEASURES AS PREDICTORS OF COMPLEX TRACKING BEHAVIOR[1]

JAMES F. PARKER, JR.
Psychological Research Associates, Inc.

AND

EDWIN A. FLEISHMAN
Yale University

THE present study investigates the relationships between ability variables and progress in learning a complex perceptual-motor skill. It represents an extension of previous research in which ability patterns at early and advanced performance levels in psychomotor tasks have been compared with a view to establishing the kinds of abilities and measures most predictive of higher levels of proficiency. These studies have implications for questions concerning the processes involved in learning complex skills as well as for problems in training and test development.

The distinction between the terms "ability" and "skill" as used here should be made explicit. The term ability refers to a more general, stable trait of the individual which may facilitate performance in a variety of different tasks. Thus, the ability of spatial-visualization may be important in such diverse activities as navigation, dentistry, and engineering. The term skill is more specific: it is task oriented. Flying an airplane is a skill, while manual dexterity and spatial-visualization are more general abilities. Of course, abilities themselves are often products of earlier learning. This view is elaborated elsewhere (Gagné & Fleishman, 1959).

It has been found that for certain classes of perceptual-motor tasks, there are progressive qualitative and quantitative changes in those abilities which contribute to proficiency as the task is learned. This has been demonstrated with a serial reaction task (Fleishman & Hempel, 1954a), a visual discrimination reaction task (Fleishman & Hempel, 1955), and a continuous pursuit task (Fleishman, 1956). Woodrow (1938) has found the same progressive changes to be involved in extended performance on printed tests. In other words, for those tasks proficiency at later stages of learning depended upon different combinations of abilities than did initial proficiency.

The present investigation differs from others in this area in two major respects. First, the task is considerably more complex than that used in previous studies. In the present instance the task roughly simulated that performed by a pilot during the attack phase of an airborne radar intercept mission. Thus, the task was more "realistic" and more demanding than the laboratory tests previously investigated. The second difference involved the total amount of learning time which was investigated. A limitation of previous studies has been the relatively short period required to achieve a high level of proficiency. In the present study subjects were trained for 17 sessions distributed over 1 to 1.5 months.

---

[1] This study was initiated in 1956 under Air Force Contract AF 41(657)-64 between Psychological Research Associates, Inc. and Air Force Personnel and Training Research Center. In 1958 cognizance of this contract was transferred to the Aero Medical Laboratory, Wright Air Development Center. Edwin A. Fleishman served as Project Monitor while a member of the Skill Components Research Laboratory of Air Force Personnel and Training Research Center. He retained his association with the project while on the faculty of Yale University. Robert A. Seibel, Bryce O. Hartman, and Marty R. Rockway served as project monitors during the latter stages of the study.

The correlation analysis and factor analysis procedures were accomplished under terms of a subcontract with Psychological Research Service, Inc. of Austin, Texas. They were directed by Benjamin Fruchter.

The tracking devices and scoring consoles were designed by Henry P. Birmingham of the Naval Research Laboratory.

Permission is granted for reproduction, translation, publication, use, and disposal in whole or in part by or for the United States Government.

The increased complexity of the present task raises the question of whether new kinds of abilities are required and whether the patterning of these abilities is different from that found with previous tasks. For example, the introduction of complexity of function may demand certain skills of integration and coordination not previously required.

A primary purpose of this study, as stated above, was to specify the abilities underlying terminal proficiency on a complex control task and to identify the manner in which the contribution of these abilities to proficiency changes during the course of learning. Another important objective was to compare the predictability of terminal performance from external measures with predictions from measures taken earlier in practice on the task itself. In a previous study (Fleishman & Hempel, 1954a) it was found that advanced performance on a psychomotor task, the Complex Coordination Test, was better predicted from a combination of printed and apparatus tests than from early scores taken on the task itself. This study presented an opportunity to test the generality of this finding in a more complex task situation.

PROCEDURES

Data Gathering

The experimental procedures followed in this study are similar to those used in previous studies in this series. Subjects were first administered a large reference battery of printed and psychomotor apparatus tests. Following this the same subjects spent a number of sessions mastering the criterion task. Performance measures were taken at the various stages of practice on this task. Correlations among scores on both the reference battery and the stages of learning of the criterion task were obtained and factor techniques applied. The factors were defined from the loadings of the reference tests. Loadings of the various stages of practice of the criterion task on the factors defined by the reference tests specify the changes in the factor pattern of this task as practice continues.

Subjects

The subjects were 203 freshman and sophomore Air Force ROTC students at the University of Maryland. No subject had had pilot training of any kind. Each subject was paid for his participation in this study.

Reference Battery

The initial testing consisted of the administration of a large battery of psychomotor apparatus and printed tests. Initially a listing was made of those factors which might be of importance in mastering a complex tracking task. This list even included such abilities as "Verbal Comprehension," inasmuch as this ability might be of importance in understanding the explanation of the requirements of the task and thus be involved in determining ultimate proficiency. The majority of the reference tests were selected in terms of their factor loadings on these abilities in previous investigations. A complete listing of the hypothesized ability factors and the loadings of the various tests on these factors is presented in Appendix A. The test battery consisted of 23 apparatus and 21 printed tests. Fifty different scores were provided by this battery.

Total time for the administration of this battery was 11 hours. This was divided into three 2-hour sessions for the administration of the printed tests and five 1-hour sessions for the administration of the apparatus tests. Two models of each apparatus test allowed for the testing of subjects in pairs. A complete listing of the tests comprising the reference battery is presented in Table 1. This table also presents reliability coefficients for each test in the battery. Appendix B presents drawings of the apparatus tests and descriptions of all tests, both printed and apparatus.

Criterion Task

Subject Task. The criterion task consisted of a tracing device constructed so as to simulate roughly the display characteristics and control requirements of an airborne radar intercept mission. The task of the subject was to maintain the target dot at the center of the oscillograph display, while at the same time nulling a sideslip indicator. That is, the subject envisioned himself to be flying the attack phase of an airborne radar intercept mission. Thus, if the target was to the right, the subject made appropriate control movements to steer the craft to the right. These movements would bring him "on target" and the dot would return to the center of the display. All turning

TABLE 1

Printed and Apparatus Tests Comprising Reference Battery

Test	Test Reliability[a]
Printed	
1. Aerial Orientation	.84
2. Complex Movements	.85
3. Coordinate Movements	.70
4. Directional Control	.85
5. Discrimination Reaction Time	.87
6. Following Directions	.75
7. Forced Landings	.88–.91
8. Formation Visualization	.76
9. General Mechanics	.88
10. Instrument Comprehension	.92
11. Mechanical Comprehension	.82
12. Pattern Comprehension	.76
13. Planning a Course	.81
14. Signal Interpretation	.77
15. Spatial Orientation	.69
16. Speed of Identification	.74
17. Stick and Rudder Orientation	.74
18. Verbal Comprehension (Vocabulary)	.90
(50) Verbal Comprehension (Background for Current Affairs)	.87
19. Visual Pursuit	.80
20. Visualization of Maneuvers	.87
Apparatus	
21. O'Connor Finger Dexterity	.76
22. Purdue Pegboard (Summation Score)	.90[b]
23. Minnesota Rate of Manipulation (Placing)	.87
24. Minnesota Rate of Manipulation (Turning)	.79
25. Two-Plate Tapping	.99
26. Precision Steadiness (Errors)	.90
27. Direction Control (Corrects)	.90–.95
28. Control Sensitivity	.94[c]
29. Ten Target Aiming (Errors)	.94
30. Ten Target Aiming (Corrects)	.91
31. Bimanual Matching	.74[c]
32. Track Tracing (Errors)	.91
33. Two-Hand Coordination	.89–.93[a]
34. Choice Reaction Time	.68[c]
35. Complex Coordinator	.85–.95
36. Pursuit Confusion (Errors)	.90
37. Pursuit Confusion (Corrects)	.88
38. Motor Judgment	.56[c]
39. Discrimination Reaction Time	.87–.96
40. Rotary Pursuit	.92–.97
41. Visual Reaction Time	.86
42. Auditory Reaction Time	.85
43. Jump Visual Reaction Time	.88
44. Jump Auditory Reaction Time	.88
45. Steadiness Aiming	.84
46. Single-Dimension Pursuit	.84–.92
47. Rate Control	.89
48. Rotary Aiming	.96
49. Rudder Control	.90–.92

---

[a] Reliabilities for those tests not computed directly from data of this study are taken from the following: Fleishman (1958b) Guilford and Lacey (1947), Roff (1951).

[b] Based on average reliability of the four subparts of this test and corrected for an increase in length by a factor of four.

[c] Computed using data of this study (split-half).

movements required coordinated action of stick and rudder controls.

Three identical tracking devices were constructed especially for purposes of this study and are shown in Figures 1 and 2. A block diagram[2] is presented in Appendix C. These devices and related scoring consoles allowed for the testing of from one to three subjects simultaneously under the control of a single test administrator.

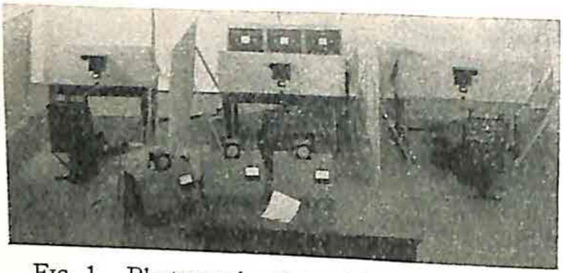


FIG. 1. Photograph of tracking facility showing three tracking devices, computer programer, and scoring console.

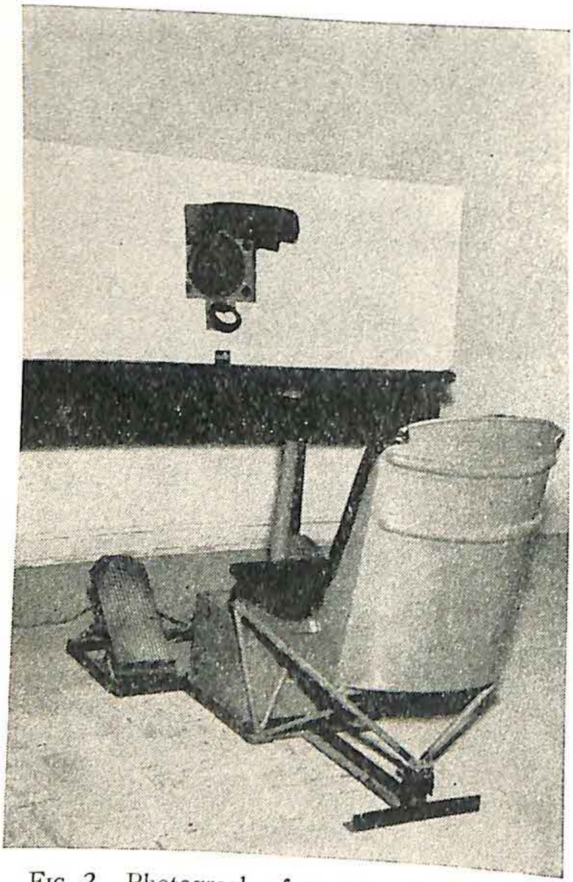


FIG. 2. Photograph of tracking device.

Each subject was given a brief indoctrination into the nature of the task and the action of the control system. He was then allowed 5 to 10 seconds of practice. Following this, through the remainder of the training program, there was no formal guidance of any sort. Each subject learned the task on his own. However, all questions of subjects related to the operation of the system were answered. Each subject was told his integrated error score following every trial.

Displays. The subject's instrument panel contained two displays. The first consisted of a target dot presented on a cathode ray oscillograph. The target course was generated by setting the equation of a swinging pendulum into an analog computer. This produced a sine wave with a frequency of approximately 6 cycles per minute in the horizontal coordinate as the target course. The rate of decay in amplitude was approximately 5% per cycle. However, the dynamic characteristics of the over-all task were such that subjects neither perceived this drop in amplitude nor the fact that the target was programed in one coordinate only. Any control imbalances resulted in dot excursions of considerably larger magnitude than those provided by target programing.

Beneath the oscillograph was an inverted 3-inch, zero-centered voltmeter termed a "sideslip indicator." This meter reading indicated to the subject a "lack of coordination" in control actions when centering the target dot. This indicator did not constitute an independent task but rather provided the subject with additional information related to the primary task.

Control System. In performing this task subjects used a standard aircraft control system involving a control stick and rudder pedals. These controls were coupled in a manner similar to those of an actual aircraft. Thus, application of right control stick pressure without proper amount of right rudder produced a sideslip to the left and a consequent left deflection on the sideslip indicator.

Control of the target dot in elevation was accomplished by forward and backward movements of the control stick. This was a pure second-order system resulting from the use of two cascaded integrators in the linkage between the control and display. Thus the acceleration of the target dot was directly proportional to stick displacement.

Control of the target dot in azimuth was accomplished by right and left movements of the control stick. This comprised a system involving acceleration control plus an exponential lag network. This dimension was mechanized using three cascaded integrators with a negative feedback loop around one. The time constant of this lag network is 1 second; that is, it requires 1 second to achieve approximately two-thirds ($1\text{-}1/e$) of the final signal resulting from a given stick displacement.

Control of the target dot in azimuth (envisoned as turning the aircraft) and centering the sideslip indicator (coordination display) were both effected by rudder pedal displacement as well as by the control stick. This rudder control of the sideslip indicator involves a simple lag network. Thus, the sideslip indicator displacement was directly proportional to rudder pedal displacement. Movement of the sideslip indicator by stick action represents a velocity control operating through two exponential lag networks.

---

[2] Complete schematics of all components are presented in *Wright Air Development Center Report* TR 59-255.

Movement of the target dot by rudder pedal displacement approximates a pure velocity control.

Criterion Scores. The following five performance measures were obtained during each testing session. Four of these were error scores; one was a time-on-target score. The error scores represent the extent of target dot (or indicator needle) displacement from the null position in one dimension summated through time.

1. Integrated absolute error score. This was recorded at the conclusion of every trial and was produced by summing algebraically the three absolute error part scores described below in accordance with this relationship:

$$T = 1/2X + 1/2Y + Z \qquad [1]$$

Where T = Integrated absolute error score
X = Absolute azimuth error
Y = Absolute elevation error
Z = Sideslip error

The double weight for the sideslip error was introduced on the basis of pretest results in which it was noted that some subjects neglected rudder pedals altogether and used only the stick to control the dot. The increased weight made a good total score more dependent upon proper rudder pedal use.

Each of the following part scores was recorded every third trial. An automatic sequencing mechanism switched from one part score to the next at the conclusion of each trial.

2. Azimuth (yaw) part score. This is a summation of the total error occurring in the horizontal axis of the oscillograph display.

3. Elevation (pitch) part score. This is a summation of the total error occurring in the vertical axis of the display.

4. Sideslip error part score. This is a measure of the lack of coordination between stick and rudder movements.

5. Time-on-target score. This score represents the time during the trial when the target dot and the sideslip indicator needle are both (simultaneously) within prescribed tolerance limits.

The time-on-target area consisted of a diamond-shaped area (a tilted square) in the center of the oscillograph. This area was not demarked for benefit of the subject. Distance from the center of the oscillograph to any corner of the square was 1 inch. Sideslip error also entered into the calculation of time-on-target. Sideslip tolerance was ± 0.5-inch needle displacement from center. Thus, if a subject approached the tolerance limit in the X axis of the scope display, a slight deflection of the sideslip indicator would throw him out of tolerance.

Data Analysis

Reference Test Scores

From each test, the obtained distributions of raw scores were transformed to normalized distributions of standard scores (stanines), each with a range from 1 to 9, a mean of 5, and a standard deviation of 2. Conversions were made so that the 9 end of the scale was always indicative of good performance.

Tracking Performance

Each of the 17 practice sessions for the criterion tracking task consisted of 21 1-minute trials, making a total of 357 trials.

The problem of deciding on the size of the time segment to be sampled from these trials involved the following considerations. It was necessary to have each segment large enough to be sufficiently reliable; at the same time, it was desired to keep the segment small enough to provide

TABLE 2

RELIABILITIES OF THE VARIOUS MEASURES OF TRACKING PERFORMANCE AT THREE STAGES DURING THE LEARNING PERIOD

	Integrated Error	Horizontal Error	Vertical Error	Sideslip Error	Time-on-Target
Early Stage (Segment 1–Segment 2)	.83	.64	.60	.63	.58
Intermediate Stage (Segment 25–Segment 26)	.87	.67	.68	.87	.78
Final Stage (Segment 50–Segment 51)	.89	.56	.67	.96	.89

Note.—All correlations based on 203 cases.

as precise as possible a definition of the learning curve. Thus, use of an entire session as a segment might be highly reliable but changes occurring during the session might be obscured.

For each session, scores of the first three trials were discarded in order to avoid problems of warm-up effect. The remaining 18 trials of each session were then subdivided into three groups of six trials each. Each group of six trials was designated "time segment" of performance. There were thus 51 time segments obtained through the full period of practice. For each of the over-all measures of performance (integrated error score and time-on-target score) one time segment represents 6 minutes of performance. For each of the part scores (elevation, azimuth, and sideslip error), a time segment represents 2 minutes of performance, a part score being recorded on every third trial only. Table 2 presents reliability coefficients for each measure of performance at an initial, intermediate, and terminal stage of learning. These coefficients represent the correlation of adjacent time segments.

Learning Curves. Figures 3 through 7 present the learning curves for the five measures of tracking performance. All error learning curves are similar except that shown in Figure 7 for sideslip (coordination) error. This error curve shows little improvement until about Segment 15. This reflects the fact that many subjects did not use the rudder pedals to any great extent during initial practice sessions. It appears that although subjects understand the requirements of this task, they will not attempt to time-share stick and rudder controls until they have attained some skill in using the stick control. The time-on-target curve differs from the others in that it is sigmoidal rather than exponential. This is because the time-on-target represents a poor measure of performance at the extremes when the subject either is continuously out of or continuously in tolerance. The error scores do not suffer in this respect.

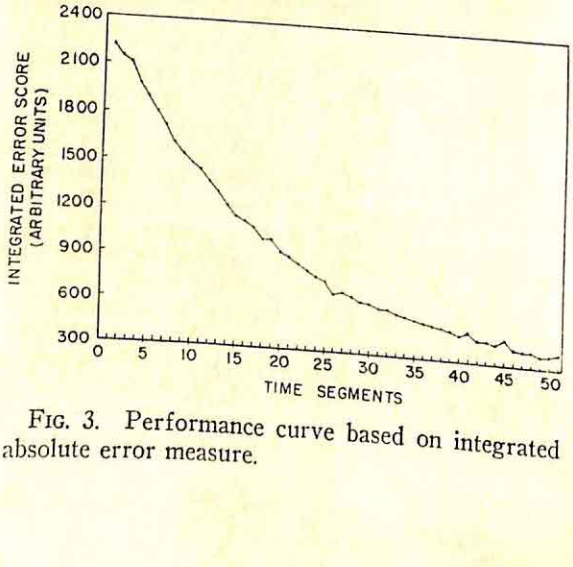


FIG. 3. Performance curve based on integrated absolute error measure.

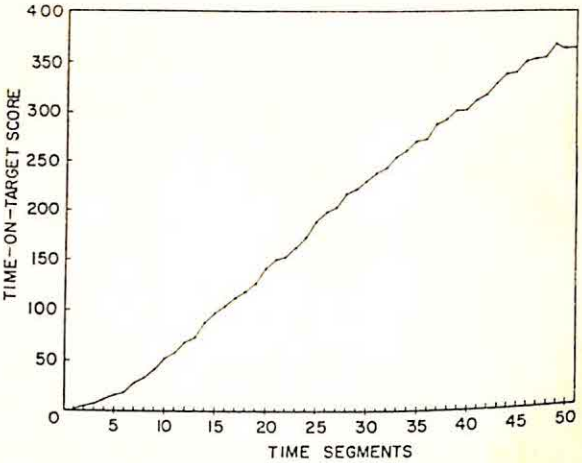


FIG. 4. Performance curve based on time-on-target measure.

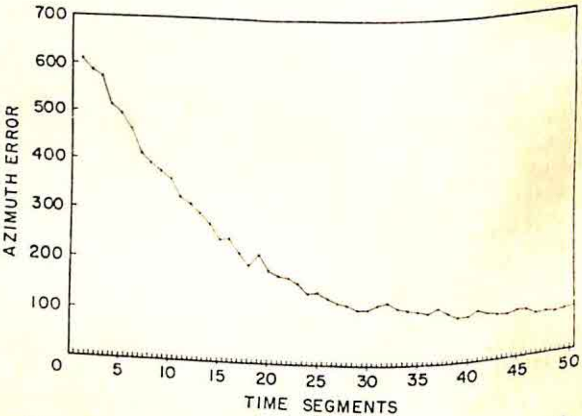


FIG. 5. Performance curve based on azimuth error measure.

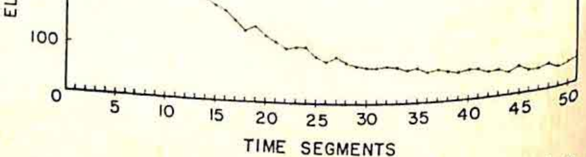


FIG. 6. Performance curve based on elevation error measure.

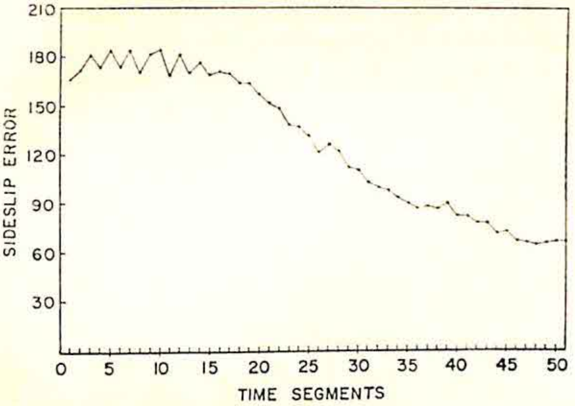


Fig. 7. Performance curve based on sideslip (coordination) error measure.

Figures 3 through 7 illustrate group performance and present relatively stable curves. There was, however, considerable variation among individuals with respect to performance on this task. Figure 8 presents performance curves based on integrated error score for three subjects. These curves were selected to illustrate the extent of individual differences both in rate of learning and in terminal proficiency. Subject A, for example, is the poorest initially but rapidly improves to a high level of proficiency. Subject B improves slowly and erratically throughout the period of practice, although initially he is better than Subject A. Subject C shows relatively little improvement at any time.

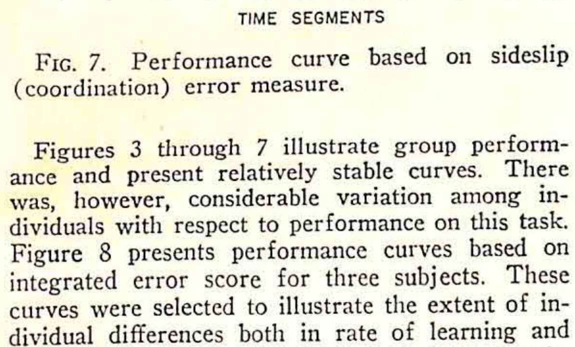


Fig. 8. Curves of three selected subjects illustrating the extent of individual variation in performance.

Selection of Practice Stages for Factor Analysis. It was decided that 10 stages of practice would be used in the subsequent factor analysis as a basis for determining the changes in underlying ability patterns during the learning of this task. The selection of the appropriate 10 stages of practice was accomplished before it was possible to compute the mean scores for the entire sample of 203 subjects. For this purpose data from a random sample of 25 subjects were chosen, and a curve using integrated error score as a measure of performance was generated. Figure 9 presents this curve and indicates the 10 stages of practice which were selected for inclusion in the final data analyses. Points were selected in a manner it was felt would maximally define critical periods during the learning process. All performance measures at each of these 10 practice stages next were converted into stanine form for further analyses.

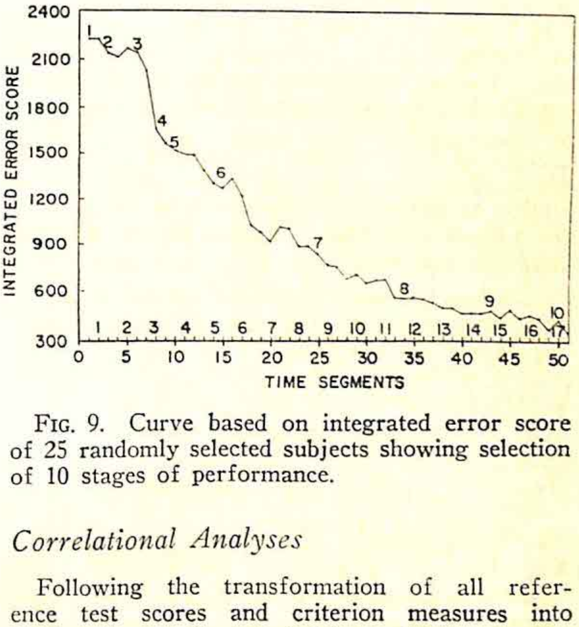


Fig. 9. Curve based on integrated error score of 25 randomly selected subjects showing selection of 10 stages of performance.

Correlational Analyses

Following the transformation of all reference test scores and criterion measures into stanine form, subsequent data analyses were accomplished through use of the ORDVAC high-speed digital computer. Appendix D presents the matrix of intercorrelations among all printed and apparatus tests of the reference battery. Appendix E presents intercorrelations between reference tests and performance measures of the tracking task. Appendix F presents intercorrelations between and within total scores and component scores taken from the task itself.

The matrix of intercorrelations of reference tests was factor analyzed using a digital computer program yielding a principal axes solution. Fifteen factors were extracted by this method. Appendix G presents the unrotated factor loadings for tests in the reference battery. Orthogonal rotations of the reference axes were made, using the graphical method of Zimmerman (1946). The correlations of performance scores (integrated absolute error) with the reference test scores next were used in projecting performance measures onto the rotated axes defined by the reference battery. The procedure used was Mosier's extension method as described by Fruchter (1954). Appendix H presents the rotated factor loadings of all reference test scores and measures of performance.

are involved, primarily (Fleishman, 1953, 1954; Fleishman & Hempel, 1955; Hempel & Fleishman, 1954).

No.	Variable	Loading
21	O'Connor Finger Dexterity	.49
22	Purdue Pegboard	.43
23	Rate of Manipulation	.36

Factor XV—Multilimb Coordination. This factor is common to tasks in which the requirement is the simultaneous use of both hands, both feet, or hands and feet. These tests have defined this factor in previous studies (Fleishman, 1958b; Fleishman & Hempel, 1956).

No.	Variable	Loading
49	Rudder Control	.40
35	Complex Coordination Test	.38
33	Two-Hand Coordination	.30

Relation of Ability Factors to Stages of Practice in Tracking

This study was concerned with determining the kinds of basic psychomotor abilities underlying tracking proficiency and with the manner in which these abilities change as a function of practice. Figure 10 illustrates the extent to which the combined abilities account for variability in tracking performance at the various stages of learning. It will be noted that only in the terminal stages do these abilities account for as much as 25% of the variance. Figure 10 clearly indicates that the set of abilities defined by the reference battery in the present study does not account for any substantial portion of the variance in tracking performance. This is in contrast to previous studies with other psychomotor tasks. For example, in the case of the Complex Coordination Test up to 75% of the variance was accounted for by independent reference ability measures (Fleishman & Hempel, 1954a); over 50% was accounted for in the Discrimination Reaction Time Test (Fleishman & Hempel, 1955); roughly 60% in the two-hand and unidimensional matching and plane control tasks (Fleishman, 1957a).

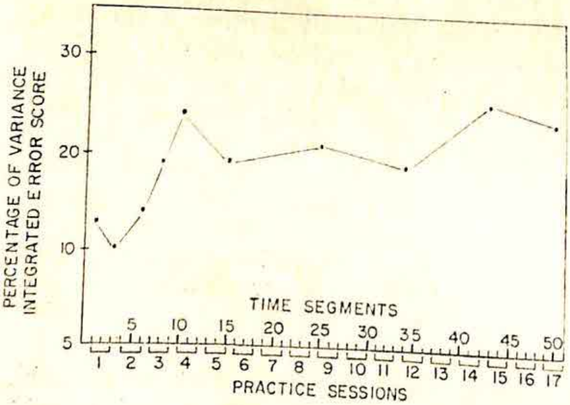


FIG. 10. Percentage of variance in tracking performance accounted for by all identified factors.

Consideration next was given to the nature of the changes occurring during practice in those abilities which do show some relation to tracking performance. Based on integrated error score as a measure of performance, only Factors I (Spatial Orientation), VII (Response Orientation), XII (Pursuit Confusion), and XV (Multilimb Coordination) had factor loadings in excess of .20 at any stage of practice. Of these, only Spatial Orientation and Multilimb Coordination show systematic changes in importance during the course of practice. Figure 11 presents curves illustrating changes in the percentage of variance accounted for by these two abilities with increased practice. Multilimb Coordination shows an increase in importance during the course of learning while Spatial Orientation appears to be important only during the middle stages.

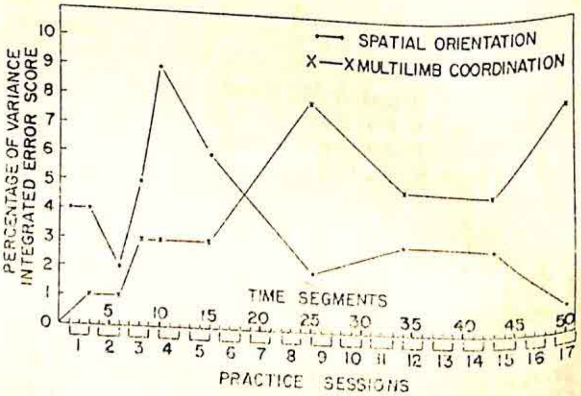


FIG. 11. Changes occurring in the importance of two ability factors with increased practice in tracking.

Consideration was given to the possibility that use of only 6 minutes of performance in the integrated error criterion measure might not have been a sufficient behavior sample and might have lowered the correlations obtained between performance and reference tests. Certainly use of an expanded measure of performance would increase the reliability of this measure even though it had previously been found to be sufficiently high (.89 at the terminal practice stage). In order to examine this possibility, several correlations between integrated error and three of the reference tests at the three stages of practice were recomputed, using a 21-minute performance measure rather than a 6-minute measure. Table 3 presents both sets of correlations. It does not appear from the changes in the correlation coefficients that any significant increase would have been obtained by using an expanded measure of performance.

Relationships among Component Performance Measures

It will be recalled that, at each of the practice stages, five measures of performance were obtained. These measures were:

1. Integrated absolute error
2. Azimuth (X) error
3. Elevation (Y) error
4. Sideslip (Z) error
5. Time-on-target (scope and sideslip simultaneously)

The integrated absolute error score is not an independent measure but is determined by the X, Y, and Z scores in accordance with the relationship described previously in Equation 1.

The above performance measures were obtained primarily with a view to answering the question: "For this type of task, how well do component performance measures predict terminal proficiency?" Table 4 presents the intercorrelations of all performance measures at an initial point in learning (first time segment) and at an intermediate point (twenty-fifth time segment).[3] The table also presents the correlations of all measures, at each of these stages, with terminal proficiency (fiftieth time segment) as measured by the integrated error score. It is apparent that initially all component measures are uncorrelated with terminal proficiency. In other words, neither initial component measures nor over-all measures allow the prediction of final performance in this task. At an intermediate point in the learning period, component measures do show some relation to the final score (X = .28, Y = .36, Z = .34) but none is sufficiently high to serve as a very useful predictor.

TABLE 3

CORRELATIONS BETWEEN SELECTED REFERENCE TESTS AND A MEASURE OF PERFORMANCE ILLUSTRATING THE EFFECT OF EXPANDING THE MEASURE OF PERFORMANCE

	Correlations	
	Previous[a]	Expanded
Aerial Orientation—Session 5	.25	.25
Discrimination Reaction Time (Printed)—Session 5	.22	.25
Complex Coordinator—Session 5	.21	.24
Aerial Orientation—Session 9	.17	.14
Discrimination Reaction Time (Printed)—Session 9	.18	.17
Complex Coordinator—Session 9	.13	.14
Aerial Orientation—Session 17	.18	.09
Discrimination Reaction Time (Printed)—Session 17	.09	.12
Complex Coordinator—Session 17	.13	.14

[a] The performance measure (integrated absolute error) entering this correlation is a 6-trial "time segment." This time segment in each case forms a part of the 21-trial session which represents the expanded measure.

---

[3] It will be recalled that the full course of practice was broken into 51 time segments.

TABLE 4

INTERCORRELATIONS AMONG COMPONENT PERFORMANCE MEASURES AND A FINAL PROFICIENCY MEASURE

	Integrated Error	X Error	Y Error	Z Error	Time-on-Target	Final Integrated Error
			TIME SEGMENT 1			
Integrated Error		.25	.15	.60	.35[a]	.04
X Error			.31	−.05	.46	.02
Y Error				−.16	.31	.08
Z Error					−.01	.12
Time-on-Target						−.02
			TIME SEGMENT 25			
Integrated Error		.65	.72	.63	.92	.46
X Error			.68	.02	.64	.28
Y Error				.27	.68	.36
Z Error					.58	.34
Time-on-Target						.48

[a] The fact that time-on-target scores are not negatively correlated with the error scores is due to the process of converting raw scores to stanine form. At this time all score distributions were arranged so that high score was indicative of good performance.

DISCUSSION

The rationale underlying this study is that there exists within the population under inquiry a finite number of abilities which are used in mastering a complex tracking task. These abilities are representative of the class of theoretical construct which is response inferred. They are response inferred in that, although the characteristics of the task remain unchanged, the performance of individuals exhibits major variation. The purpose of this study was to identify these abilities in terms of relations among the various performance measures (reference battery). Variability in performance at different stages of mastering the tracking task then could be accounted for in terms of the identified abilities. The validity of such ability identification is attested to by the factorial invariance found in previous studies using different reference tests (see Appendix A). The present study confirms this invariance using different subjects and a different combination of tests.

If tracking performance in fact can be accounted for in terms of a limited number of measurable human abilities, the results will have implication for such areas as human engineering design of equipment, personnel measurement and selection, and the identification of an efficient set of learning principles designed to improve training in tracking procedures.

Initially it might be well to indicate the rationale behind this study and the manner in which this study relates to the broad area of tracking research. Adams (1961) indicates that there exist two major kinds of tracking research. Engineering psychologists have used the theory of closed-loop servomechanisms as a model for investigating the man-machine tracking system. Adams points out the advantages and disadvantages of this orientation and notes that it results in research oriented toward *task variables*. Thus, system input-output relations are discussed in terms of time and frequency characteristics and are described through output errors resulting as a function of changes in system parameters. Such changes might be performed on the input signal, control system dynamics, or the control-error display relationships.

General experimental psychology, on the other hand, has used traditional behavior theory as its model and the tracking system becomes a means of eliciting desired behavior classes. Such research is oriented toward *procedural variables* related to training, retention, stress, fatigue, motivation, etc. The major deficiency in this type of research is that apparently any task is considered appropriate if it elicits the desired response class. Little consideration is given to possible interactions between task and procedural variables.

The present investigation appears to fall somewhere between these two types of research. It is concerned with task variables to the extent of insuring a complex criterion task and defining the dimensions of complexity. These are a two-dimensional display system, control characteristics varying from zero-order to second-order, and requirement for dual-error nulling. Task variables were not used as independent variables, however. This study was not concerned with procedural variables as such, yet the results have implications for such areas as learning theory and training.

Identification of Perceptual-Motor Abilities

One major end product of this study is found in the identification of the ability factors measured by the tests within the reference battery. Forty-three printed and apparatus tests were administered to these subjects. (Fifty scores appeared due to use of subparts and correct and error scores as separate measures of performance.) Although the majority of these tests were chosen on the basis of their factor loadings in previous investigations, a number of new tests were introduced. Results of the factor analysis of these test intercorrelations serve further to define the basic components of human perceptual-motor abilities.

Changes in Ability Patterns with Practice

The general finding in investigations of this type (Fleishman, 1956, 1959) is that changes in the ability pattern underlying proficiency in the criterion task are quite systematic and progressive during the learning period. For the tracking task of this study, however, Figure 11 indicates that only Factor I (Spatial Orientation) and Factor XV (Multilimb Coordination) show a systematic relation to tracking performance during the course of practice. Apparently, early in practice the ability to interpret the visual display characteristics is of some importance. Later the ability to balance and to coordinate the activity of hands and feet in manipulating the control stick and rudder pedals plays some part in determining the proficiency of the subject. This is in keeping with the observation during the testing program that most subjects experienced extreme difficulty in effecting coordinated movements during the early part of the training program.

The Multilimb Coordination factor begins its major increase in importance at approximately the fifteenth time segment (Figure 11). This is roughly the same period when the sideslip (Z) error score, which reflects coordinated hand-foot activity, begins to show considerable improvement (Figure 7). There are thus two independent analyses in support of each other. An identified ability factor and a measure of component task activity both indicate the importance of coordination between limbs in determining proficiency at a given point in the practice schedule.

Predictability of Tracking Performance

One basic problem for consideration with any complex activity is whether terminal proficiency is better predicted by a set of external measures or by scores taken from initial practice periods in the activity itself. With the present tracking task there is no correlation between initial tracking performance and terminal performance. (The correlation between first-segment integrated error scores and fiftieth, or final, integrated error scores is .04). Based on the first time segment intercorrelations shown in Table 4, a multiple correlation analysis was performed. Results of this analysis indicate the extent to which terminal performance is predictable through use of the five performance measures taken during the initial practice session. The obtained multiple correlation coefficient is .17, which is not statistically significant. In other words, scores obtained during initial practice periods in the activity itself do not predict terminal proficiency. It is interesting to note that even at the twenty-fifth time segment, when the subject has had approximately 4 hours of practice in the device, the correlation of integrated error score with terminal proficiency (terminal integrated error score) has risen only to .46.

In opposition to the above findings there are six of the reference tests which show correlations with final tracking performance of .26 or higher. The intercorrelations among these tests are presented in Table 5. The obtained multiple correlation using these tests as predictors of terminal proficiency is .43, which is significant at the .01 level.

These results indicate that early component and total scores taken from the task itself do not predict advanced proficiency levels until an intermediate stage of practice, and these are not high. If one must choose between intratask and independent measures, selected independent measures will give a better prediction of advanced tracking proficiency.

TABLE 5

RESULTS OF MULTIPLE CORRELATION ANALYSIS PREDICTING TERMINAL TRACKING PERFORMANCE FROM SIX EXTERNAL MEASURES

Variable	11	36	37	39	40	K
7[a]	.35	.25	.29	.37	.14	.26
11		.27	.19	.21	.21	.26
36			.48	.26	.31	.26
37				.31	.31	.26
39					.29	.30
40						.27

$R = .43^*$

[a] These numbers correspond to the designation of the following variables during the machine data analyses:
Var. 7 = Forced Landings
Var. 11 = Mechanical Comp.
Var. 36 = Pursuit Conf. (Err.)
Var. 37 = Pursuit Conf. (Corr.)
Var. 39 = Discrimination Reaction Time (Appar.)
Var. 40 = Rotary Pursuit
Var. K = Segment 50 (Terminal) Integrated Error Score

* Significant at the .01 level ($N = 203$).

Human Performance in Tracking

The over-all operation of tracking a moving target can be examined in terms of component responses required of the human being. Adams (1961) has classified the component responses in tracking behavior into observing, predicting, and motor responses. These are defined below and are followed by a listing of those factors identified in this study which should be important to each response. In some instances certain factors appear to emphasize elements of more than one response class.

Observing Response. The sensing of display parameters including both input signal and error signal is by the observing response. The observing response appears to involve head and/or eye movements to direct the visual receptors to spatially separated stimuli, and the discrimination of stimulus change. The following identified factors would seem to be important for the observing response:

Factor I.	Spatial Orientation
Factor IX.	Perceptual Speed
Factor X.	Visualization

Prediction Response. The prediction reponse involves predicting the regularities inherent in the input signal and the effect of control movements upon the error signal. This implies that the subject has learned the dynamics of the control system and the interactions between control movements and display changes. Poulton (1957) refers to such prediction as perceptual anticipation. Perceptual anticipation involves prediction of nature and size of muscular contractions required, prediction of duration of response movement, and prediction of future position of target at time of completion of response movement. For the criterion task of this study the following factors would appear to be important:

Factor VII.	Response Orientation
Factor X.	Visualization
Factor XI.	Integration

The Rate Control factor, which was originally hypothesized, would fit in this category. However, this factor was not identified in the present study.

Motor Response. The motor responses are those activities involved in control system (stick and rudder) manipulations. It appears that the following abilities identified within this study would be of importance to such motor responses:

Factor II.	Control Precision
Factor III.	Speed of Arm Movement
Factor IV.	Manual Dexterity
Factor VII.	Response Orientation
Factor VIII.	Arm-Hand Steadiness
Factor XI.	Integration
Factor XII.	Pursuit Confusion
Factor XIV.	Finger Dexterity
Factor XV.	Multilimb Coordination

Identification of Abilities Underlying Tracking Proficiency

The preceding analysis would indicate that a successful tracking response is dependent upon a number of component responses. These component responses would, in turn, involve certain basic "abilities." The major objective of this study was to determine the extent to which a set of abilities as identified through use of the reference battery of tests accounts for variability in tracking performance at various stages in the practice schedule. The possibility existed that all of the factors listed in the preceding section would be important. As it turned out, this was not the case. As seen in Figure 10, at no stage of practice did these abilities, taken in their entirety, account for over 25% of the variance in tracking. As seen in Figure 11, only two of the identified abilities show any systematic relation to tracking performance. These are Spatial Orientation, listed under Observing Responses, and Multilimb Coordination, listed under Motor Responses. However, it should be noted that the contribution of the observing response factor decreases through practice while the motor response factor increases in importance.

The major question now becomes: "Why were these abilities not more effective in accounting for variability in tracking performance?" The need for some answer becomes more obvious when it is realized that in a previous study (Fleishman & Hempel, 1955) using a different criterion task (the Discrimination Reaction Time Test—DRT), over 60% of criterion task performance could be accounted for at some practice stages using the same kinds of abilities as in this instance. This 60% is as opposed to less than 25% within this study.

A possible answer may rest both with the nature of the tests within the reference battery and the nature of the criterion task. Considering the criterion task first, it has been noted that this tracking task was quite complex due to the dual display, the second-order control system characteristics, and the requirement for dual-error nulling. Another way of viewing this task might be that the subject was required to maintain a highly complicated and unstable system in a state of equilibrium. With tasks investigated previously such as the DRT, Rotary Pursuit, and the Complex Coordination Test, it appears reasonable that different abilities might be required for mastery. For example, with the DRT the subject observes a visual stimulus which differs on successive presentations only in terms of four possible configurations. The subject reacts by pushing one of four toggle switches as quickly as possible. For this task it appears logical that such abilities as Spatial Orientation, Reaction Time, and Rate of Arm Movement would be important.

The fact that a similar set of abilities did not account for a substantial amount of the variance in performance in the present instance may be explained if the task is examined in light of the three required reponse areas. The *observing response* in this task requires that the subject time-share between two separate features of the over-all display. The *prediction response* requires that the subject develop the capability of predicting target position at some future time. When it is considered that the display actually is an error signal produced by improper control movements, it is apparent that prediction of target position would require a rather complex ability. This is particularly true since two dimensions of the control system involved second-order controls. Thus the prediction of target position might involve such hypothesized abilities as the ability to perform double- and higher-order differentiations while compensating for built-in control system lag effects. Certainly proficiency in this task would require a different set of abilities than would be found with the DRT or the other tasks investigated previously.

The particular selection of tests for the reference battery used in this instance represents an attempt to continue with the same type of tests used so successfully in previous investigations employing different criterion measures. The lack of success of this comprehensive battery of tests in defining a set of abilities which would relate to tracking proficiency warrants a closer look at the tests selected. It appears that a set of abilities are important with this type of criterion task for which appropriate predictive tests have not been developed. Further work is required for the construction of such tests. Thus, if the criterion task is known to require time-sharing activities, the reference battery should contain tests designed to measure an hypothesized general time-sharing ability. There are no known reference tests which might be used to measure such an ability.

In short, while the reference tests used in this study reaffirmed a number of basic human perceptual-motor abilities, these abilities could not be related to any considerable extent to tracking performance. A complex tracking activity apparently requires abilities for which appropriate reference measures are not available at present. These abilities, further, do not fall within the area customarily called "Motor Abilities." Purely motor abilities do not determine individual differences in advanced tracking proficiency. The important abilities appear to fall within the areas of "Observing Responses" and, propably to a greater extent, "Prediction Responses." Further studies, obtaining basic measures within these areas, should be able to contribute additional information concerning the nature of a complex tracking activity.

SUMMARY AND CONCLUSIONS

This investigation was designed to provide information concerning the kinds of abilities which best predict performance in a complex tracking task. Consideration was given to the manner in which the underlying ability pattern exhibits change as a function of increasing proficiency in the task. The relation between early component performance measures and terminal proficiency in the tracking task also was examined.

A battery of reference tests containing both apparatus and printed tests of known factorial structure was administered to 203 subjects. Following this each subject spent a total of 17 sessions mastering the tracking task. Scores on the reference tests and various learning levels of the tracking task were analyzed through factor analysis techniques. Fifteen ability factors were identified and were related to the different stages of practice in tracking.

The following conclusions are drawn from the results of this study:

1. The kinds of perceptual-motor abilities which previously have been found to be related to performance on laboratory criterion tasks do not account for any substantial amount of the variance in performance with the present task. It is hypothesized that the complexity and particular dynamics of the tracking task used in this study require different kinds of abilities. Such abilities as Rate of Arm Movement apparently are not as important to the mastery of complex tracking activities as with laboratory psychomotor tasks. In tracking, the important determiners of proficiency may be found in those abilities used in the analysis of target motion and in the prediction of appropriate response movements, rather than in abilities related directly to control manipulation.

2. Initial component and total scores taken from the tracking task itself do not predict advanced proficiency levels. If one must choose between intratask and independent measures, the independent test measures give a better prediction of advanced tracking proficiency.

REFERENCES

Adams, J. A. Human tracking behavior. *Psychol. Bull.*, 1961, in press.

Fleishman, E. A. Testing for psychomotor abilities by means of apparatus tests. *Psychol. Bull.*, 1953, **50**, 241-242.

Fleishman, E. A. A factorial study of psychomotor abilities. *USAF Personnel Train. Res. Cent. res. Rep.*, 1954, No. 54-15.

Fleishman, E. A. Predicting advanced levels of proficiency in psychomotor skills. In G. Finch & F. Cameron (Eds.), *Symposium on Air Force human engineering, personnel, and training research.* (NAS-NRC Pub. 455) Washington, D.C.: NAS-NRC, 1956. Pp. 142-151.

Fleishman, E. A. A comparative study of aptitude patterns in unskilled and skilled psychomotor performance. *J. appl. Psychol.*, 1957, **41**, 263-272. (a)

Fleishman, E. A. Factor structure in relation to task difficulty in psychomotor performance. *Educ. psychol. Measmt.*, 1957, **17**, 522-532. (b)

Fleishman, E. A. An analysis of positioning movements and static reactions. *J. exp. Psychol.*, 1958, **55**, 13-24. (a)

Fleishman, E. A. Dimensional analysis of movement reactions. *J. exp. Psychol.*, 1958, **55**, 438-453. (b)

Fleishman, E. A. Le pronostic des niveaux élèves d'aptitude dans taches complexes. Bull. *Ass. Int. Psychol. Appl.*, 1959, **8**, 26-43.

Fleishman, E. A., & Hempel, W. E. Changes in factor structure of a complex psychomotor test as a function of practice. *Psychometrika*, 1954, **19**, 239-252. (a)

Fleishman, E. A., & Hempel, W. E. Factorial analysis of complex psychomotor performance. *USAF Personnel Train. Res. Cent. res. Rep.*, 1954, No. 54-12. (b)

Fleishman, E. A., & Hempel, W. E., The relation between abilities and improvement with practice in a visual discrimination reaction task. *J. exp. Psychol.*, 1955, **49**, 301-312.

Fleishman, E. A., & Hempel, W. E. Factorial analysis of complex psychomotor performance and related skills. *J. appl. Psychol.*, 1956, **40**, 96-104.

French, J. W. The description of aptitude and achievement tests in terms of rotated factors. *Psychometr. Monogr.*, 1951, No. 5.

Fruchter, B. *Introduction to factor analysis.* New York: Van Nostrand, 1954.

Gagné, R. M., & Fleishman, E. A. *Psychology and human performance.* New York: Holt, 1959.

Guilford, J. P., & Lacey, J. I. (Eds.) Printed classification tests. *AAF Aviat. Psychol. Program res. Rep.*, 1947, No. 5.

Hempel, W. E., & Fleishman, E. A. A factor analysis of physical proficiency and manipulative skill. *USAF Personnel Train. Res. Cent. res. Rep.*, 1954, No. 54-340.

Michael, W. B., Guilford, J. P., Fruchter, B., & Zimmerman, W. S. The description of spatial-visualization abilities. *Educ. psychol. Measmt.*, 1957, **17**, 185-199.

Poulton, E. C. On prediction in skilled movements. *Psychol. Bull.*, 1957, **54**, 467-478

Roff, M. F. Personnel and classification procedures: Spatial tests. USAF Sch. Aviat. Med. *proj. Rep.*, 1951, No. 21-29-002.

Woodrow, H. The relation between abilities and improvement with practice. *J. educ. Psychol.*, 1938, **29**, 215-230.

Zimmerman, W. S. A simple graphical method for orthogonal rotation of axes. *Psychometrika*, 1946, **11**, 51-55.

(Received March 31, 1960)

APPENDIX A

TESTS WHICH MEASURE BASIC PSYCHOMOTOR ABILITIES AS DETERMINED IN PREVIOUS STUDIES

FACTOR	APPARATUS TEST	(LOADING, REFERENCE)[a]	PRINTED TEST	(LOADING, REFERENCE)[a]
1. Psychomotor Coor. I (Fine sensitivity) (1, 2, 3)[a]	1. Rotary Pursuit	(.48, 1)(.53, 2)(.47, 3)		
	2. Complex Coor. Test	(.48, 1)(.36, 2)(.50, 3)		
	3. Rudder Control	(.45, 2)		
	4. Pursuit Confusion	(.48, 3)		
2. Psychomotor Coor. II (Gross sensitivity; multilimb movements) (3, 4)	1. Rudder Control	(.56, 3)		
	2. Plane Control	(.50, 3)		
	3. Multi-dimen. Pursuit	(.38, 3)		
	4. Complex Coor.	(.37, 3)(.41, 4)		
	5. Rotary Pursuit	(.42, 4)(.08, 3)		
3. Speed of Arm Movement (1)	1. Reaction Time	(.54, 1)		
	2. Plane Control	(.33, 1)		
	3. Disc. Reac. Time	(.25, 1)		
	4. Rotary Pursuit	(.34, 1)		
	5. Rate of Movement	(.48, 1)		
4. Spatial Orientation (Stimulus interpretation) (3)	1. Controls Orientation	(.46, 3)	1. Direction Cont.	(.39, 3)
	2. Drift Correction	(.40, 3)	2. Instru. Comp.	(.46, 1)(.37, 4)
	3. Direction Control	(.38, 3)		
	4. Disc. Reac. Time	(.37, 3)		
5. Response Orientation (Response choice) (3)	1. Disc. Reac. Time	(.50, 3)	1. Signal Disc.	(.52, 3)
	2. Complex Mult. Reac.	(.41, 3)	2. Disc. Reac. Time	(.50, 3)
6. Perceptual Speed (1)			1. Visual Pursuit	(.50, 1)
			2. Speed of Ident.	(.47, 1)
			3. Spatial Orien.	(.45, 1)
			4. Disc. Pursuit	(.55, 3)
			5. Coordination	(.42, 3)
7. Visualization (1)	1. Direction Control	(.44, 3)	1. Pattern Control	(.60, 1)
	2. Controls Orien.	(.36, 3)	2. Mech. Principles	(.41, 1)
			3. Decoding	(.36, 1)

8. *Mechanical Experience* *(1)*			1. General Mechs.	(.62, 1)(.70, 4)
			2. Mech. Principles	(.49, 1)(.44, 4)
			3. Tool Functions	(.69, 4)
9. Numerical Facility (1)			1. Numer. Oper's II	(.66, 1)
			2. Dial & Table Reading	(.60, 1)
10. Psychomotor Speed (1, 2)			1. Log Book Accuracy	(.55, 1)(.59, 2)
			2. Speed of Marking	(.50, 1)
			3. Speed of Ident.	(.37, 1)
			4. Marking Accuracy	(.46, 2)
11. Wrist-Finger Speed (2)			1. Large Tapping	(.74, 2)
			2. Med. Tapping	(.74, 2)
			3. Pursuit Aiming—1	(.50, 2)
12. Finger (Fine dexterity) (2, 4, 5)	1. Purdue Peg Bd Both Hands	(.61, 2)		
	2. Purdue Peg Bd Left Hand	(.58, 2)		
	3. Purdue Peg Bd-Assm.	(.55, 2)(.35, 4)		
	4. Santa Anna Dext. Test	(.46, 4)		
13. Rate of Arm Movement (Probably same as Factor 2) (2)	1. 10-Target Aiming-Cor.	(.66, 2)		
	2. Two Plate Tapping	(.54, 2)		
	3. Rotary Aiming	(.46, 2)(.57, 4)		
	4. Rate of Movement	(.45, 4)		
14. Aiming			1. Pursuit Aiming—II	(.68, 2)
			2. Pursuit Aiming—I	(.63, 2)
			3. Aiming	(.63, 2)
			4. Irreg. Dotting Pursuit	(.82, 5)
			5. Speed of Sq. Marking	(.71, 5)
15. Arm-Hand Steadiness (2, 5)	1. Track Tracing	(.61, 2)		
	2. Steadiness-Aiming	(.60, 2)		
	3. Precision-Steadiness	(.50, 2)		
	4. Track Steadiness	(.56, 5)		
16. Reaction Time (2, 4)	1. Visual Reac. Time	(.73, 2)(.72, 4)		
	2. Auditory Reac. Time	(.68, 2)(.68, 4)		
	3. Jump Vis-Reac. Time	(.73, 4)		
	4. Jump Aud-Reac. Time	(.70, 4)		

APPENDIX A—Continued

FACTOR	APPARATUS TEST	(LOADING, REFERENCE)[a]	PRINTED TEST	(LOADING, REFERENCE)[a]
17. Manual Dexterity (2, 3, 5)	1. Santa Anna Finger Dext.	(.47, 2)(.38, 3)		
	2. 10-Target Aiming-Err.	(.43, 2)		
	3. Min. Rate of Manipulation (Turning)	(.38, 2)		
	4. Rotary Pursuit	(.35, 3)		
	5. DRT	(.34, 3)		
	6. Marble Board	(.51, 5)		
18. Postural Discrimination (2)	1. Post. Disc.—Angular	(.55, 2)		
	2. Post. Disc.—Vertical	(.48, 2)		
19. Integration (Coordination of hands and feet) (3)	1. Multi-dimen. Pursuit	(.36, 3)		
	2. Complex Coor.	(.36, 3)		
20. Rate Control (3)	1. Rate Control	(.58, 3)		
	2. Single-dimen. Pursuit meter	(.55, 3)		
	3. Compensatory Balance	(.39, 3)		
21. Verbal Comprehension (4)			1. Word Knowledge	(.78, 4)
			2. Background for Current Affairs	(.77, 4)
			3. Pattern Comp.	(.46, 4)

---

[a] References as indicated by numbers within parentheses:

1. FLEISHMAN, E. A., & HEMPEL, W. E. Changes in factor structure of a complex psychomotor test as a function of practice. *Psychometrika*, 1954, **19**, 239-252.
2. FLEISHMAN, E. A. Dimensional analysis of psychomotor abilities. *J. exp. Psychol.*, 1954, **48**, 437-454.
3. FLEISHMAN, E. A., & HEMPEL, W. E. Factorial analysis of complex psychomotor performance. *USAF Personnel Train. Res. Cent. res., Rep.*, 1954, No. 54-12.
4. FLEISHMAN, E. A., & HEMPEL, W. E. The relation between abilities and improvement with practice in a visual discrimination reaction test. *J. exp. Psychol.*, 1955, **49**, 301-312.
5. HEMPEL, W. E., & FLEISHMAN, E. A. A factor analysis of physical proficiency and manipulative skill. *USAF Personnel Train. Res. Cent. res., Rep.*, 1954, No. 54-34.

APPENDIX B

DESCRIPTION OF TESTS IN REFERENCE BATTERY

Printed Tests

1. *Aerial Orientation.* The subject is required to match the cockpit view of the horizon that corresponds to the airplane from which this view would be seen.

2. *Complex Movements.* Each problem presents a vertical line and above it a horizontal line. On each of these lines is a circle and a dot. Below the lines is a grid of 49 squares, one of which contains an "S." The examinee indicates in the grid from the letter S the direction required to "move" the dots on both lines simultaneously into their respective circles.

3. *Coordinate Movements.* Each problem presents a series of arrows which vary in curvature and represent different directions of movement. The thickness of each arrow represents distance of movement. To the left of these arrows are a dot and circle. The subject selects the one symbol (arrow) that will move the dot inside the circle.

4. *Directional Control.* Each problem presents nine boxes side-by-side with a target circle above the middle box. Also pictured is an arrow whose head indicates direction and whose length indicates distance of movement. The subject indicates from which box it must be launched to have its head reach the target.

5. *Printed Discrimination Reaction Time.* Each item simulates a setting of the apparatus test of the same name except responses are made by pencil in one of four slots on answer sheets. The slots are arranged in the same up-down, left-right pattern as are the toggle switches on the apparatus version.

6. *Following Directions.* Two panels of typical aircraft display instruments are presented to the subject. As the subject reads the text he underlines certain words depending upon relations among the instruments and upon directions appearing within the text. The test is designed to measure ability to work rapidly under the pressure of complex directions.

7. *Forced Landings.* Aircraft are depicted at certain altitudes above a number of boxes representing landing fields. Arrows are also shown on the display representing wind direction and magnitude and up or down draft currents. The subject must assess the situation and select the best landing field in terms of prescribed criteria.

8. *Formation Visualization.* The subject is presented with the top and side views of silhouettes of aircraft flying in formation. He then must select the correct alternative from among four front views of this same formation flight.

9. *General Mechanics.* Printed items require practical mechanical information dealing with the use of operation of familiar mechanical methods and devices.

10. *Instrument Comprehension.* For each item which presents views of cockpit instruments the subject must determine the proper position and orientation of an airplane.

11. *Mechanical Comprehension.* The subject is presented drawings of mechanical devices. He then must answer a series of questions concerning the operation of each device.

12. *Pattern Comprehension.* A series of drawings is presented which require visualization of relationships between components of solids and their unfolded, flat projections.

13. *Planning A Course.* The subject is presented a drawing of a maze-like arrangement of boxes. Starting at one end the subject must trace his way through the boxes, responding to coded instructions as to direction of travel which are presented within the various boxes.

14. *Signal Interpretation.* The subject is presented with a picture of a row of ships, each ship having an arrow beneath it. The subject traces from ship to ship in accordance with the pattern of flags being flown by each ship. Each problem continues until the subject returns to a ship used previously.

15. *Spatial Orientation.* The subject is presented with a large aerial photograph. This is followed by several small photographs, each presenting an expanded view of one section of the large photograph. The subject must connect each small photograph with the proper position within the large photograph.

16. *Speed of Identification.* Pictorial items are presented in which the silhouette of an object must be identified when it is rotated and imbedded in a group of similar silhouettes.

17. *Stick and Rudder Orientation.* A series of pictures is presented to the subject showing changing views through the windscreen of an aircraft. The subject is required to recognize the corresponding change in the position of the aircraft and to indicate the movements of stick and rudder required to produce such a change.

18. *Verbal Comprehension (Vocabulary).* This is a typical vocabulary test requiring the matching of words and definitions.

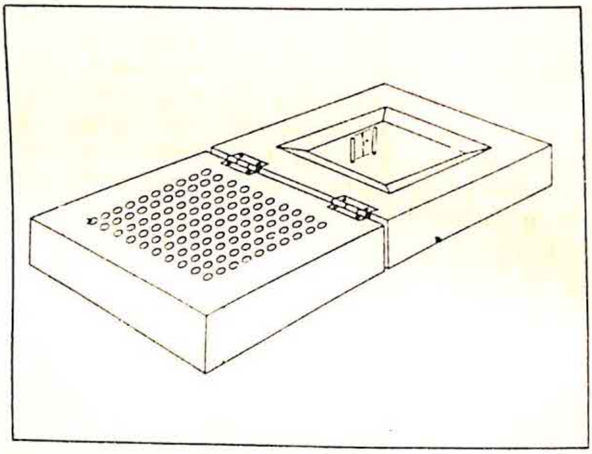
(50.) *Verbal Comprehension (Background for Current Affairs).* This is an informational test covering current, recent, and historical events.

19. *Visual Pursuit.* From a series of mazes or irregularly curved lines the task is to trace each line visually from its beginning to its proper termination point.

20. *Visualization of Maneuvers.* A picture of an aircraft is presented. This is followed by three descriptions of changes in the aircraft position. The subject then selects from among five alternatives that picture of the airplane which represents its final position following the described maneuvers.

Apparatus Tests

21. *O'Connor Finger Dexterity.* The subject is required to pick up three small pins at a time from a tray of pins with the preferred hand and place them three at a time in a small hole. He must fill a series of small holes in this manner as fast as possible. Score is the number of pins placed. One 5-minute trial.



22. *Purdue Pegboard (Summation Score).*

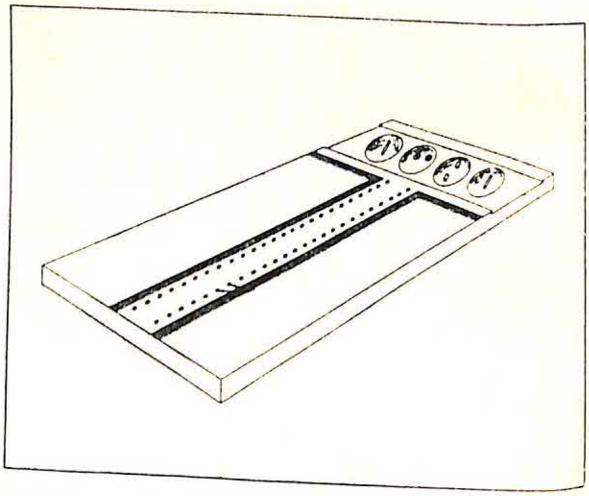
Right hand: The subject is required to place a number of small pegs individually in a series of small holes as rapidly as possible with the right hand. Score is the number of pegs placed. One 30-second trial.

Left hand: Same except with the left hand.

Both hands: The subject is required to pick up two pins at a time, one with each hand from different trays and place them simultaneously in two different holes. Score is the number of pegs placed. One 30-second trial.

Assembly: The subject is required to make as many completed peg-washer-collar-washer assemblies as possible in the time allowed. Score is the number of assembly components completed. One 30-second trial.

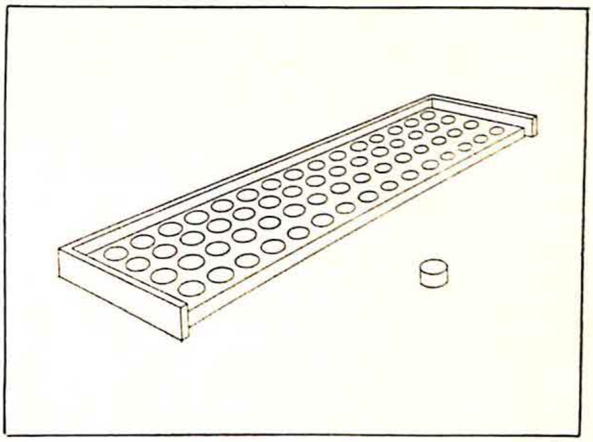
A summation of the four scores was used in the data analysis.



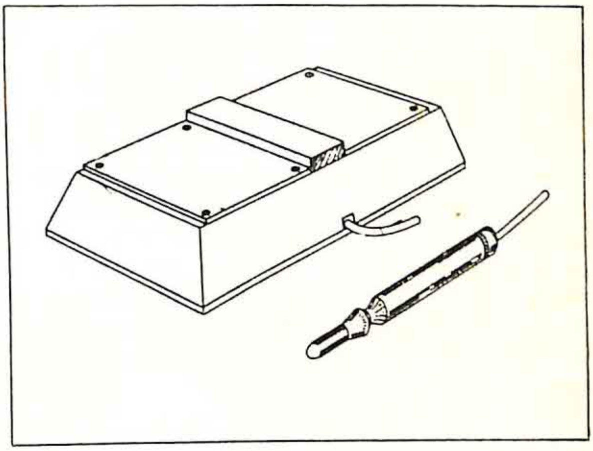
23. *Minnesota Rate of Manipulation.*

Placing: The subject is required to place 60 cylindrical blocks in the proper holes as rapidly as possible. The score is the number of blocks placed. Two 35-second trials.

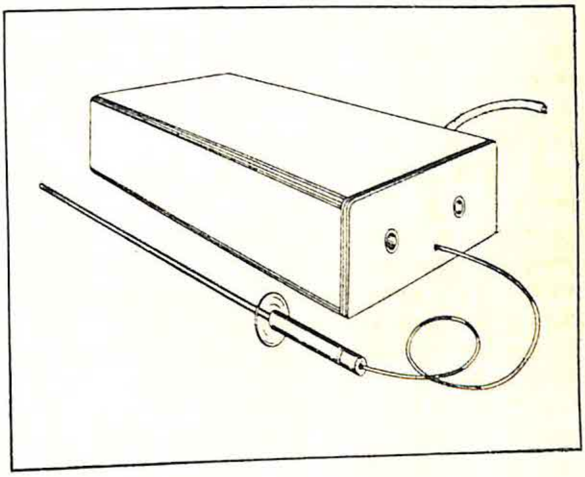
Turning: The subject is required to remove the blocks from the holes with one hand, turn them over with the other hand, and replace them in the same holes, moving from block to block as rapidly as possible. Score is the number of blocks turned. Two 40-second trials.



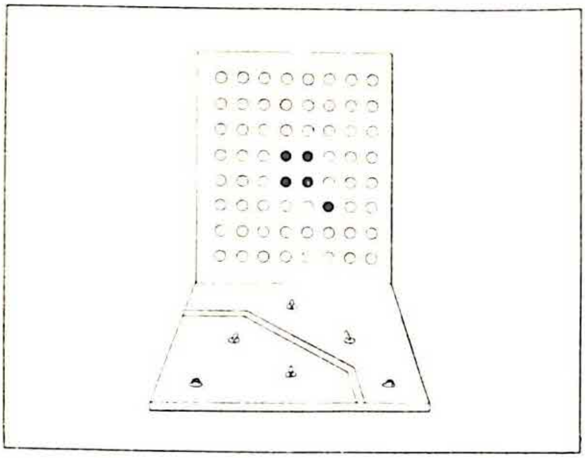
24. *Two-Plate Tapping.* The subject is required to strike two adjacent metal plates with a stylus as rapidly as possible. He strikes the plates successively; that is, first one and then the other, making as many taps as possible on the plates in the time allowed. The number of taps is recorded on counters. Three 30-second trials.



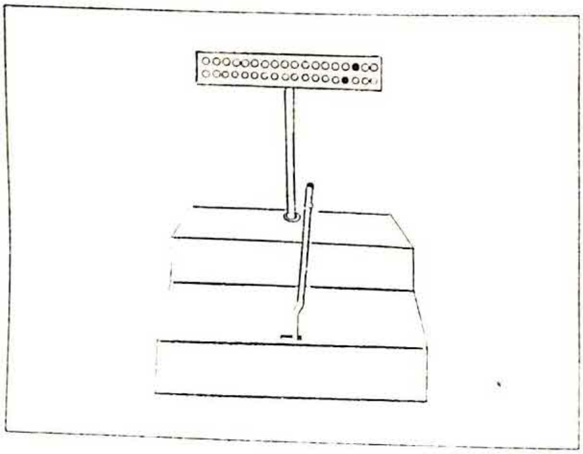
25. *Precision Steadiness.* The subject is seated before a long, rectangular box-like apparatus containing two openings. Each opening is the entrance to a straight passageway which the subject must negotiate with a long stylus. He moves the stylus forward at slightly below shoulder height and at arm's length. He must move the stylus slowly and steadily away from his body trying not to hit the sides of the cylindrical passage. As he reaches the end of the passage he strikes a contact point and withdraws the stylus, again trying to avoid hitting any part of the passageway. He then negotiates the second passageway. Two complete negotiations constitute a trial. Counters record the number of contacts and clocks record the number of times in contact. Six trials, no time limit.



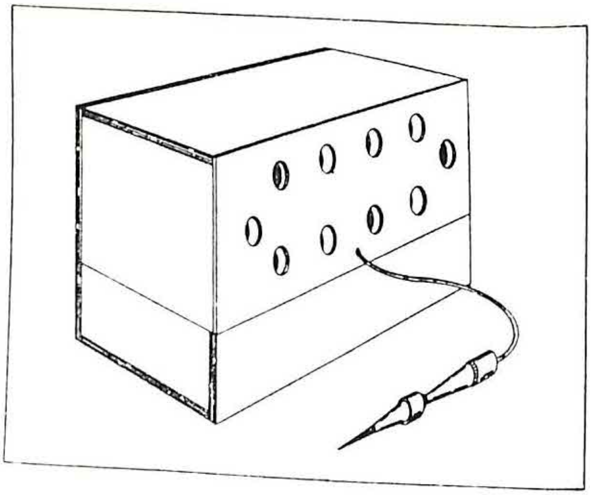
26. *Direction Control.* Patterns of lights appear in the form of a cross (to indicate the "heading" of an airplane). The pattern may appear in different parts of the display panel and may represent a small or large cross, depending upon the direction of heading. The examinee must use both hands concurrently in manipulating the proper combination of switches or switches and buttons. A new aircraft appears upon proper manipulation. Four 2-minute trials.



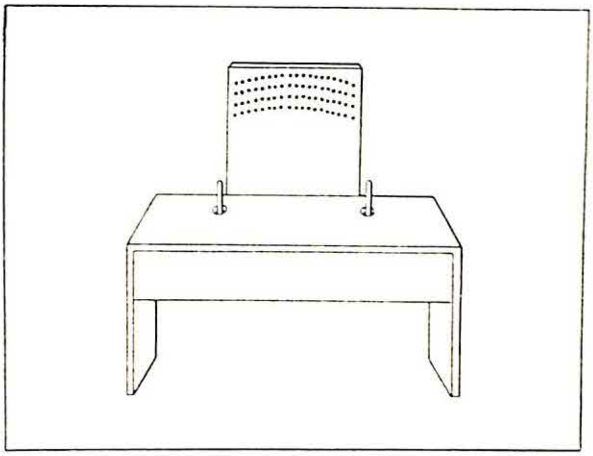
27. *Control Sensitivity.* The subject is required to match the position of a red light with a corresponding green light. The position of the lighted green light is controlled by a highly sensitive control stick. A slight movement of the stick to the right displaces the lighted green light to the right and, similarly, for left movements. When the subject has matched the two lights and held this position for .5 second, the red light moves to a new position and the subject again proceeds to match it. Score is the number of completed matches in four 1-minute trials.

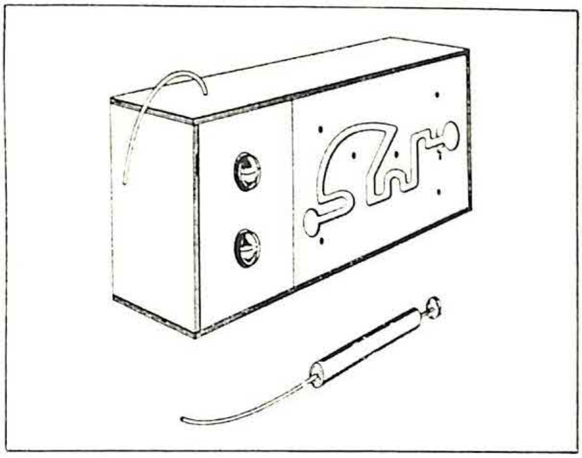


28. *Ten Target Aiming.* The subject is seated before an upright panel containing 10 holes arranged at equal intervals in an elliptoid pattern. Behind each hole can be seen a circular target. These targets vary in size from hole to hole. The subject is required to strike at these targets with a stylus moving from target to target around the pattern of targets in a clockwise direction. He is instructed that both speed and accuracy count and that he must try to hit as many targets as possible, striking only once at a target. Error counts are recorded each time the subject strikes the outside of each hole or inside around the target area. Correct counts are scored for hits within the target area in each hole. Six 30-second trials.



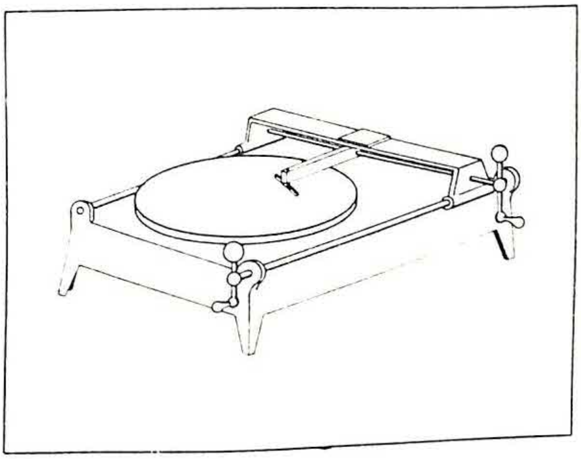
29. *Bimanual Matching.* A red row of stimulus lights has a parallel row of green lights above and below it. The subject must match lights simultaneously in each row of green lights with the red lights through movements of two handles. The movements of each hand sometimes correspond and sometimes are antagonistic to each other and bear a complex relationship to the position of the stimulus lights from setting to setting. Score is the number of simultaneous matchings made and held for .5 second during each of two 2-minute trials.





30. *Track Tracing.* The subject is required to negotiate an irregular slot pattern with a T-shaped stylus held at arm's length. Score is the number of errors (contacts with the back, top, and sides of slot) during four attempts. No time limit.

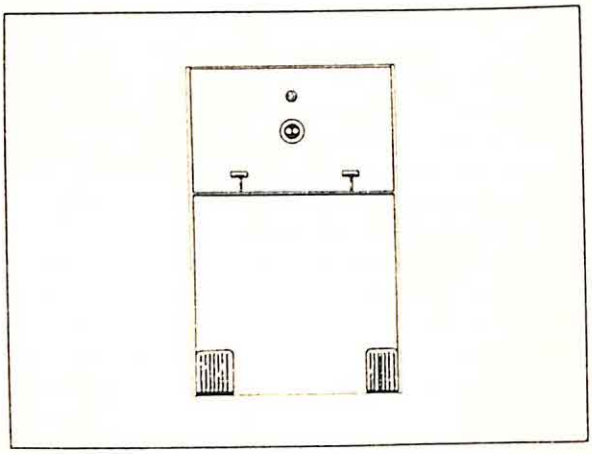
31. *Two-Hand Coordination.* The subject attempts to keep a target follower on a small target disc as the target moves irregularly and at varying rates. Movement of the target follower to the right and left is controlled by one lathe-type handle, movement to and from is controlled by the other. Consequently, simultaneous rotation of both handles moves the follower in any resultant direction. Score is the total time-on-target during four 1-minute trials.



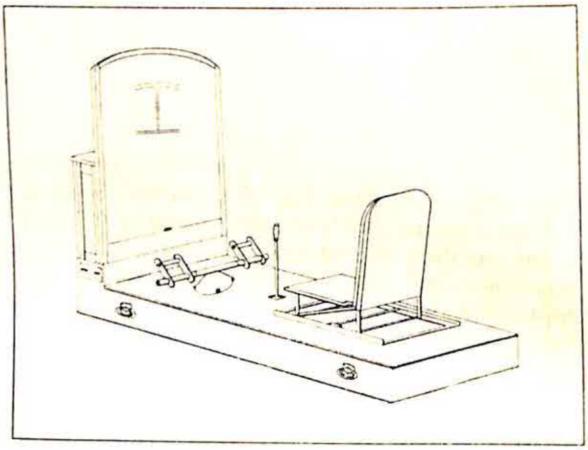
32. *Choice Reaction Time.* The subject is confronted with a panel containing a central light. He has two levers, one for the right hand and one for the left hand, and two foot pedals. He responds appropriately to any of four stimuli as follows:

red light—right lever forward
green light—left lever back
buzzer—right pedal depressed
bell—left pedal depressed

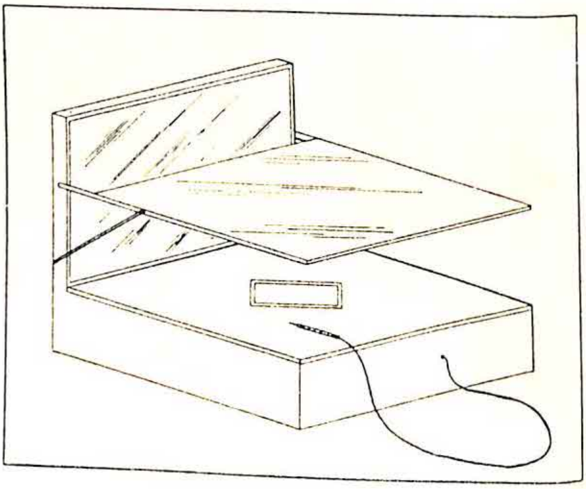
The subject is allowed 3 seconds in which to respond. The central light extinguishes when the correct response is made.



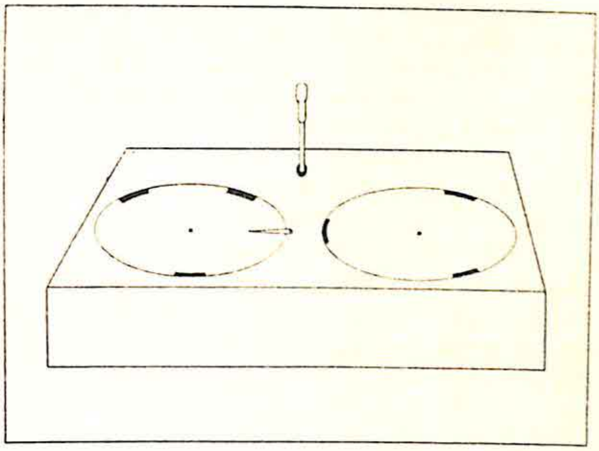
33. *Complex Coordination Test.* The subject is required to make motor adjustments of an airplane-type stick and rudder in response to successively presented visual patterns. A correct response (movement of stick and rudder controls to proper positions) is not accomplished until both the hands and feet have completed and maintained the appropriate adjustment. A new pattern appears as each correct response is completed. Score is the number of correct responses completed in a given test period. Four 2-minute trials.



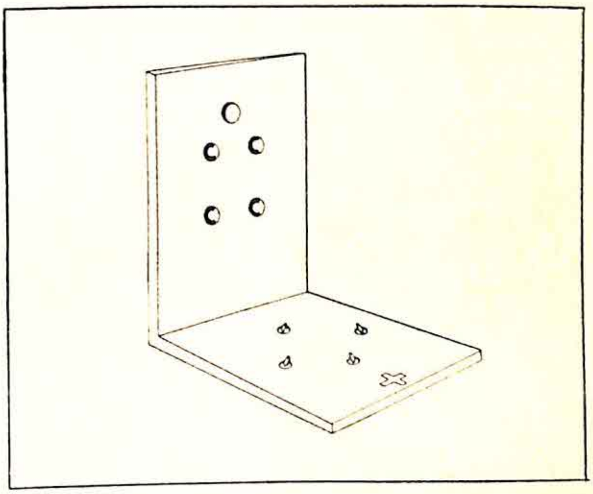
34. *Pursuit Confusion.* The subject attempts to keep a stylus on a variable speed target as it moves through a diamond-shaped slot. The task is complicated by the fact that the entire target area is visible only by mirror vision. Score (corrects) is the time-on-target during the six 1-minute trials. Error score is the amount of time the subject is in contact with the sides of the slot.



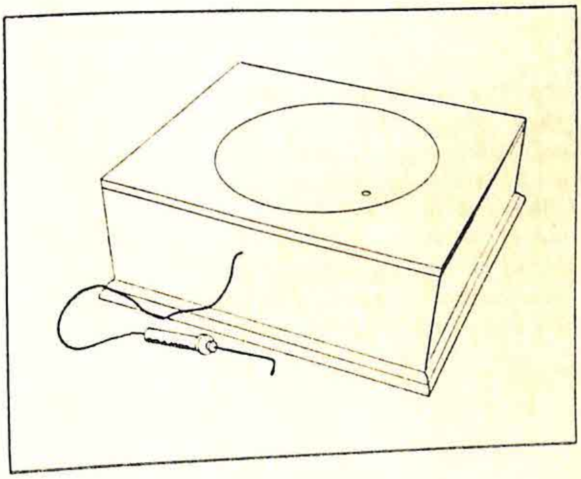
35. *Motor Judgment.* The subject is confronted by two adjacent discs rotating at a constant speed. Each disc has black and white sections on its perimeter. Between these discs is a pointer whose speed of rotation the subject can control. A forward movement of this stick slows the pointer and a backward movement speeds up the pointer. The subject cannot stop the rotation of the pointer completely and he can exert no control over the two rotating discs. The subject is required to manipulate the control stick so as to make as many revolutions of the pointer as possible without crossing the black areas on the rotating discs. To do this properly he must integrate his estimates of speed of each disc, the pointer, and his own control movements. Score is the ratio of the number of errors to the number of pointer revolutions during four 1-minute trials.



36. *Discrimination Reaction Time.* The subject manipulates one of four toggle switches as quickly as possible in response to a series of visual stimulus patterns differing from one another with respect to the spatial arrangement of their component parts, e.g., position of a lighted red lamp relative to a lighted green lamp. Score is the accumulated time of response for four series, each series containing 20 reactions for each stimulus pattern (subject must respond within 3 seconds).



37. *Rotary Pursuit.* The subject attempts to keep a prod stylus in contact with a small metallic target set in a rapidly revolving phonograph-type disc. Score is the total time-on-target during five 20-second trials.



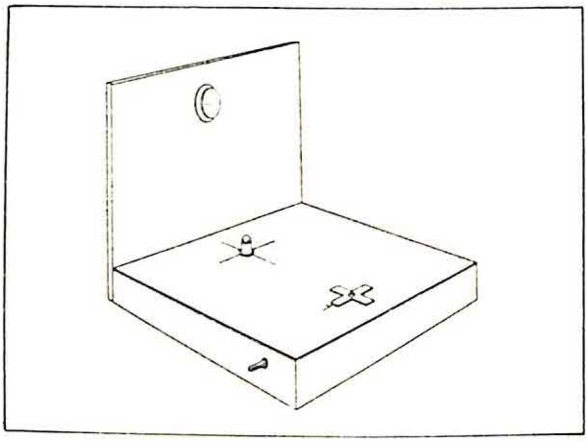
38. *Reaction Time Test.*

Visual: The subject keeps his finger on a button, depressing it as rapidly as possible in response to a single amber light before him. A click provides him with a ready signal before each light stimulus is presented, with the fore period (between click and light) varying in a random order from .5 to 1.5 seconds. Score is accumulated reaction time for a series of 20 reactions.

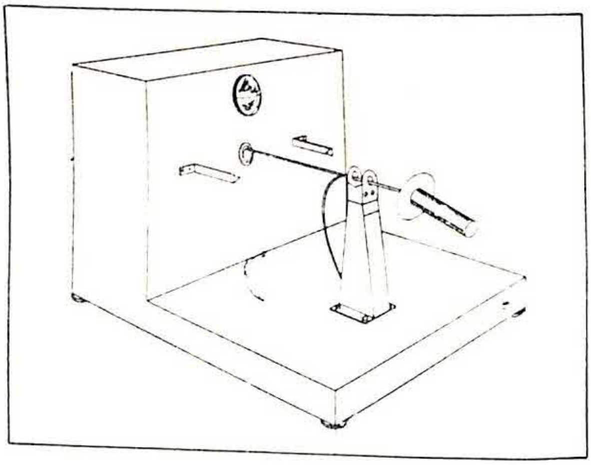
Auditory: Same procedure except that the response is to a buzzer.

Jump Visual: Same procedure except that the subject does not keep his finger on the response button but keeps it on a cross 6 inches from the response button. He must move his hand to the button as rapidly as possible as each light stimulus appears.

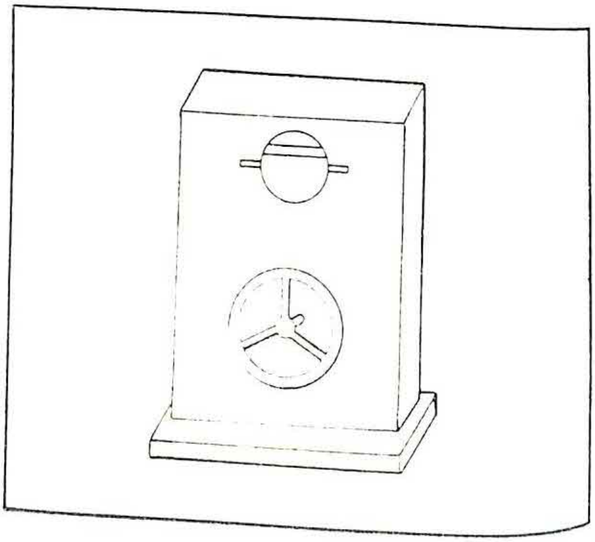
Jump Auditory: Same except response is to buzzer.



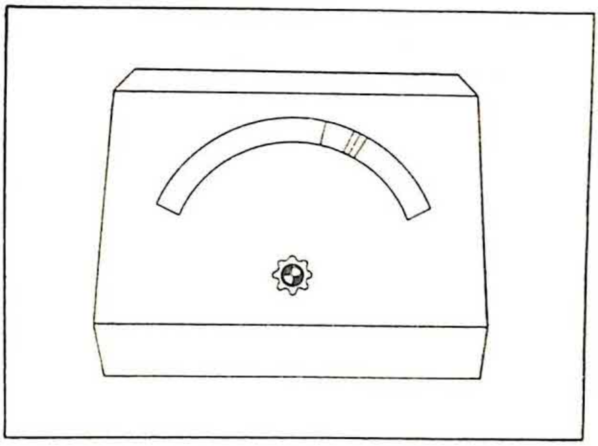
39. *Steadiness Aiming.* The subject must keep a delicately balanced stylus centered in a small hole. Any contact with the sides of the hole is recorded on a clock. The score is the number of seconds in contact. Six 40-second trials.



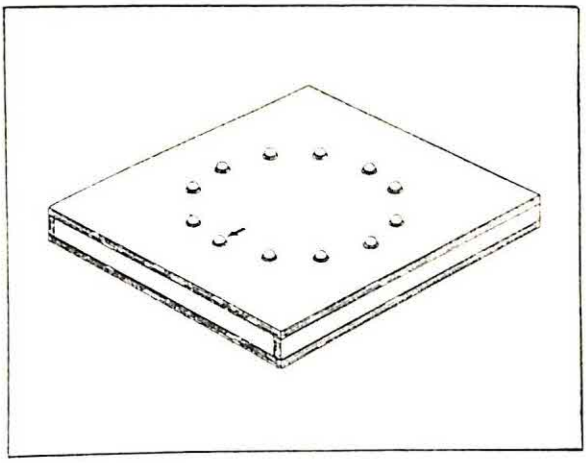
40. *Single Dimension Pursuit.* The subject makes compensatory adjustments (in and out movements) of a control wheel in order to keep a horizontal line in a null position as it deviates from center in irregular fashion. The control wheel is damped pneumatically, introducing a lag into the system. Score is the time the horizontal line is held in a null position during the four 1-minute trials.



41. *Rate Control.* A vertical target line moves back and forth across a curved scale with frequent changes in direction and rate of movement. The subject attempts to keep a thin pointer in coincidence with this line by adjustive rotary manipulations of a knob control. Score is the total time the pointer and target line are in coincidence during eight 1-minute trials.



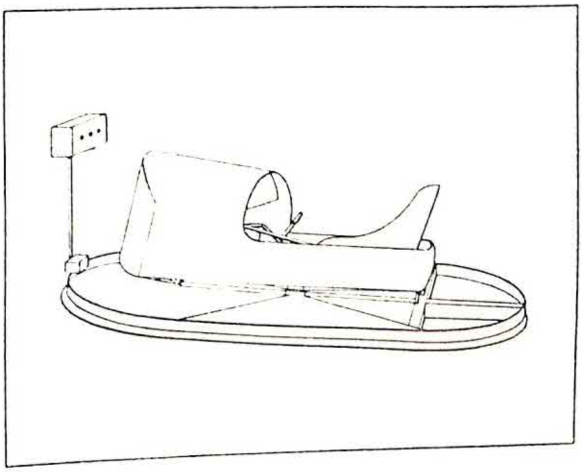
42. *Rotary Aiming.* The task is to strike at a series of buttons arranged in a circular pattern on a horizontal panel, going from one button to the next as rapidly as possible. Score is the number of strikes in four 30-second trials.



43. *Rudder Control.*

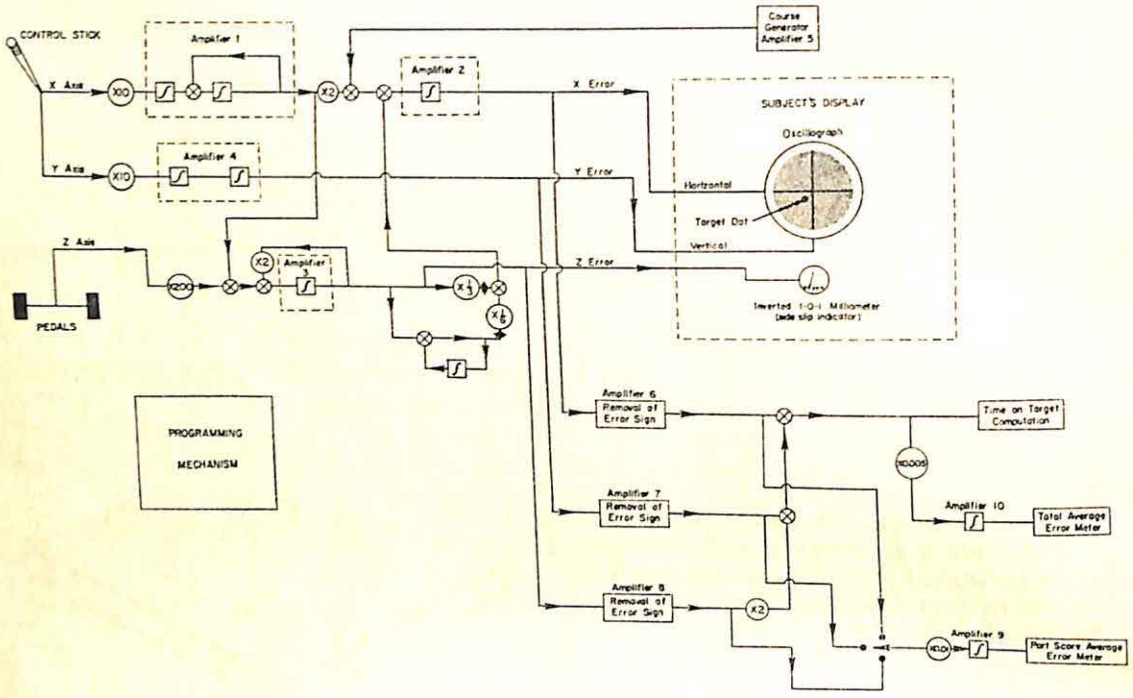
Center Target: The subject sits in a mock airplane cockpit which he attempts to keep aligned with one of three target lights as they come on in front of him. His own weight throws the seat off-balance unless he applies and maintains proper correction by means of foot pedals. In this condition only the center target is used. Score is the total time the cockpit is held aligned with the center light during three 30-second trials.

Triple Target: Same procedure as above except the subject must utilize appropriate pedal control to shift the cockpit from one light to another as these come on at random intervals. Score is the total time the cockpit is lined-up with the proper light during three 112-second trials.



APPENDIX C

BLOCK DIAGRAM OF TRACKING DEVICE



SYSTEM BLOCK DIAGRAM

APPENDIX D

INTERCORRELATIONS, MEANS, AND STANDARD DEVIATIONS AMONG REFERENCE TESTS

VARIABLES	NO.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	Mean	SDs	No.
Aerial Orientation	1		21	40	32	24	30	26	45	45	54	43	38	34	30	22	19	55	28	38	54	02	10	-04	07	11	17	40	08	02	12	26	20	30	07	26	30	16	11	21	24	-03	07	-07	04	06	05	27	-02	31	16	5.03	1.93	1
Complex Movements	2			36	30	37	24	22	44	27	39	29	41	35	33	21	23	37	10	35	37	19	23	07	16	08	15	36	22	10	05	27	27	22	10	25	26	25	07	36	24	05	07	04	11	01	14	18	02	25	09	5.21	2.10	2
Coordinate Movements	3				40	32	38	40	43	29	40	37	40	35	41	27	25	46	17	39	36	05	14	-00	05	13	18	39	07	-00	06	23	19	27	06	20	28	17	07	29	14	03	12	09	06	-01	13	13	04	22	19	4.96	1.92	3
Directional Control	4					27	18	43	28	17	41	28	27	25	34	21	30	33	15	27	28	15	25	09	13	12	25	30	11	01	-04	29	15	27	06	25	26	21	06	27	27	04	06	03	04	-02	07	12	12	17	01	5.03	1.94	4
Discrimination Reaction Time (Printed)	5						20	28	25	21	39	17	27	25	21	25	36	19	10	28	27	12	21	14	18	22	13	36	25	-07	16	27	24	24	22	30	30	29	06	36	23	14	03	10	-28	-32	31	24	32	33	14	5.06	1.98	5
Following Directions	6							21	29	-11	01	16	16	14	23	25	36	32	26	35	34	-02	03	-12	-04	-06	21	20	09	06	-00	15	08	13	-02	13	11	10	-10	17	06	-04	-05	01	01	-07	05	11	-06	11	23	5.00	1.94	6
Forced Landings	7								31	29	37	35	30	29	34	16	32	35	18	32	38	13	22	15	10	17	18	29	23	-01	04	27	26	24	15	30	25	29	01	37	14	-03	14	02	14	12	03	16	-01	14	03	4.97	1.91	7
Formation Visualization	8									51	56	55	67	35	33	35	24	57	25	38	58	13	11	-06	01	00	20	41	11	11	-07	22	26	21	09	32	26	16	-03	31	17	-09	00	-04	12	12	11	21	-15	23	11	5.01	1.93	8
General Mechanics	9										49	63	53	28	22	26	15	45	24	33	40	14	13	-02	10	09	25	34	28	17	02	32	30	29	03	37	38	37	10	23	29	-13	05	-04	12	14	10	26	-06	37	11	4.97	1.97	9
Instrument Comprehension	10											45	50	39	44	37	35	56	28	37	48	04	16	-04	12	06	19	42	22	14	-01	31	25	28	20	36	35	27	09	36	23	-06	-02	-05	05	07	12	31	-03	26	13	5.05	1.96	10
Mechanical Comprehension	11												50	33	32	17	16	51	29	30	46	10	04	-03	00	-01	19	38	11	05	04	23	17	29	03	30	27	19	02	21	21	-11	04	-00	12	13	11	19	-11	26	20	5.24	2.02	11
Pattern Comprehension	12													35	28	22	19	52	23	39	49	21	23	04	11	06	25	44	09	11	05	28	27	22	05	19	32	28	00	35	22	-11	05	-01	06	08	04	13	-06	29	08	5.05	1.77	12
Planning A Course	13														48	25	25	38	09	41	31	-04	11	00	13	02	20	36	19	04	09	21	16	18	04	30	11	21	01	24	06	02	11	02	02	07	00	21	-04	21	23	4.99	2.00	13
Signal Interpretation	14															21	13	37	11	35	28	02	08	-09	02	09	13	25	10	03	-07	26	14	22	14	20	23	22	10	21	00	-03	04	-05	07	-01	06	22	-13	26	03	5.04	1.96	14
Spatial Orientation	15																32	32	-10	30	32	12	30	17	29	17	21	21	19	-08	13	20	19	16	13	31	18	18	04	18	08	-06	06	12	16	03	17	25	02	12	05	5.02	1.95	15
Speed of Identification	16																	-16	18	39	23	19	20	22	15	11	25	16	26	-04	-12	16	18	11	19	29	13	09	-01	21	17	-02	07	08	06	01	07	17	21	17	21	4.96	1.94	16
Stick and Rudder Orientation	17																		31	38	56	02	10	-12	06	-02	19	38	12	06	07	25	22	28	-02	31	33	24	01	30	25	-14	01	-02	07	12	16	25	-08	29	07	5.16	1.99	17
Vocabulary	18																			12	26	00	-08	-12	-06	-07	-02	11	04	-19	-20	-05	01	01	05	02	-10	05	15	-10	-03	01	-04	-01	-03	01	12	05	-01	06	51	4.96	1.89	18
Visual Pursuit	19																				44	29	23	05	13	07	20	33	17	05	13	34	27	33	03	38	32	28	12	27	14	-08	-00	06	07	-03	14	33	08	31	-01	4.84	1.83	19
Visualization of Manuevers	20																					-01	16	-03	03	07	13	41	16	-02	06	29	17	23	07	30	29	23	02	31	22	-20	-12	00	10	-03	14	25	-14	26	05	5.03	1.90	20
Finger Dexterity	21																						37	26	27	16	20	11	06	06	16	07	18	16	08	16	14	15	08	20	13	03	06	10	12	06	13	08	20	21	-08	5.06	1.93	21
Purdue Pegboard	22																							48	43	33	20	19	14	-00	18	26	32	15	10	20	23	21	13	23	23	-03	07	18	14	-04	03	21	19	15	-06	4.95	1.91	22
Rate of Manipulation (Placing)	23																								44	37	07	11	18	-04	20	03	09	03	13	08	06	09	14	16	21	09	06	15	14	01	08	04	23	10	-09	4.94	1.92	23
Rate of Manipulation (Turning)	24																									39	12	09	30	-06	17	21	12	22	06	22	09	26	07	20	23	00	11	08	11	02	15	12	25	24	-02	5.02	1.91	24
Two Plate Tapping	25																										15	10	28	-11	25	24	22	15	22	18	07	11	12	17	27	03	06	07	12	00	11	13	22	29	-09	4.99	2.00	25
Precision Steadiness (Errors)	26																											19	19	-01	18	29	39	14	10	26	09	16	13	22	26	11	12	15	12	15	08	25	09	37	-02	4.96	1.96	26
Direction Control	27																												20	11	01	23	22	35	13	35	25	22	12	32	25	-00	-06	-06	04	-01	16	15	00	26	-07	5.01	1.95	27
Control Sensitivity	28																													-10	29	26	16	20	25	48	16	29	09	30	21	09	12	20	21	10	16	25	08	32	-03	5.04	1.94	28
Ten Target Aiming (Errors)	29																														-55	06	09	-03	-12	-02	08	-00	10	-07	-01	-11	-08	-18	-07	-02	-09	-03	-01	00	-05	4.99	1.92	29
Ten Target Aiming (Correct)	30																															14	13	12	16	23	-01	05	-01	12	20	18	11	22	12	14	10	14	20	21	-05	5.06	1.89	30
Bimanual Matching	31																																30	37	16	38	36	30	13	17	34	-04	-02	01	11	07	07	31	06	43	-16	4.97	1.94	31
Track Tracing (Errors)	32																																	25	22	34	23	25	14	22	27	07	05	04	18	27	05	27	19	36	07	4.96	1.93	32
Two-Hand Coordination	33																																		13	45	31	36	19	28	31	-07	06	06	21	06	24	27	04	48	-10	5.02	1.97	33
Choice Reaction Time	34																																			25	-01	08	03	22	-01	20	07	11	17	20	-07	13	15	15	08	5.01	1.82	34
Complex Coordinator	35																																				26	33	13	31	25	08	10	13	16	17	21	41	12	44	-06	5.00	1.90	35
Pursuit Confusion (Errors)	36																																					48	20	26	31	-14	-04	-01	02	06	15	22	-04	30	08	4.97	1.98	36
Pursuit Confusion (Correct)	37																																						23	31	31	-14	04	07	08	09	24	22	-01	39	-10	5.02	1.96	37
Motor Judgment	38																																							08	26	-03	02	-02	-04	03	03	15	02	25	12	5.06	1.95	38
Discrimination Reaction Time	39																																								29	11	12	28	26	18	11	27	01	24	04	5.04	1.99	39
Rotary Pursuit	40																																									-06	-11	03	06	07	17	23	16	44	-12	4.93	1.99	40
Visual Reaction Time	41																																										38	24	26	19	-07	08	16	-04	-00	4.66	1.16	41
Auditory Reaction Time	42																																											23	32	07	01	23	21	-01	-14	5.03	1.43	42
Jump Visual Reaction Time	43																																												48	02	13	16	15	04	-05	4.33	1.72	43
Jump Auditory Reaction Time	44																																													19	19	29	17	20	01	5.53	1.71	44
Steadiness Aiming	45																																														03	14	-01	16	01	4.96	1.94	45
Single Dimension Pursuit	46																																															25	07	19	-00	5.00	1.96	46
Rate Control	47																																																13	36	-04	4.92	1.95	47
Rotary Aiming	48																																																	09	-03	5.11	1.81	48
Rudder Control	49																																																		-10	4.99	1.95	49
Current Affairs	50																																																			4.93	1.87	50

APPENDIX E

CORRELATIONS OF REFERENCE MEASURES WITH TRACKING PERFORMANCE MEASURES

TRACKING MEASURES

Reference Measures[a]	Integrated Error										Azimuth Error (X)										Elevation Error (Y)										Sideslip Error (Z)										Time-On-Target									
	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
1	16	18	23	22	29	25	17	24	24	18	08	-05	17	17	20	16	17	20	17	18	05	07	16	27	23	20	16	10	20	12	18	11	-01	05	10	03	14	11	17	10	11	09	16	18	22	23	20	23	21	23
2	07	01	09	12	15	17	12	16	20	11	08	04	18	14	15	18	16	05	11	05	17	09	05	13	23	17	10	11	14	03	05	-03	-07	-05	-02	-04	05	13	12	05	06	12	12	10	15	21	17	16	18	12
3	17	10	12	18	17	25	16	25	28	16	03	05	13	15	14	23	09	18	12	05	18	12	12	24	19	22	19	16	21	02	10	03	05	-13	01	07	12	18	25	16	07	13	15	18	17	27	18	25	25	22
4	12	15	11	14	24	12	14	15	19	17	06	01	10	10	18	18	14	03	09	14	09	14	04	13	23	17	17	09	18	13	11	-02	-03	-03	00	-02	07	15	20	06	01	07	10	09	21	14	16	16	18	17
5	06	10	17	27	24	22	18	19	20	09	07	04	21	22	21	19	15	03	14	02	26	17	21	31	23	25	19	03	11	04	01	-05	-13	-07	-07	-01	14	16	20	15	13	15	23	34	26	23	22	22	20	16
6	17	13	06	10	11	10	12	12	12	05	09	05	02	04	-02	10	08	06	05	03	10	04	00	07	06	14	03	06	12	02	20	12	07	04	04	-01	11	10	03	02	10	10	06	10	03	10	10	15	08	10
7	07	10	12	14	26	20	20	17	22	26	10	07	03	08	15	13	19	14	07	09	03	09	07	10	14	19	19	09	18	10	03	-05	-02	02	05	06	12	14	26	18	04	08	10	13	22	23	21	21	23	25
8	10	13	14	19	21	18	16	20	20	11	00	02	16	13	14	16	15	04	02	07	-01	09	13	18	16	15	14	17	10	06	15	07	01	03	05	03	11	14	19	07	03	13	13	14	16	22	17	24	18	17
9	17	16	14	16	20	17	10	14	16	16	13	01	11	05	12	15	10	06	09	11	06	07	13	09	16	13	05	15	11	07	17	03	05	03	04	02	12	05	15	03	12	16	12	14	16	21	15	15	13	18
10	22	25	24	28	36	31	20	22	20	19	15	07	20	17	31	24	23	10	09	15	16	17	18	24	31	25	17	14	03	12	17	01	-03	05	03	13	12	20	21	13	17	19	24	27	36	32	22	26	21	23
11	17	16	19	19	26	22	18	21	22	26	12	03	16	11	19	17	11	09	09	21	03	13	19	18	25	20	09	17	10	14	13	04	-02	-04	11	04	19	10	19	18	13	14	16	16	21	25	20	22	19	28
12	13	13	15	11	18	20	05	11	13	09	05	03	12	07	15	16	08	02	-03	-03	03	09	10	10	17	16	11	15	03	05	10	-03	02	-01	-01	06	03	01	12	05	03	13	11	11	14	21	09	15	03	12
13	13	14	14	15	24	20	11	14	17	09	07	00	02	11	07	19	13	08	09	09	14	04	07	13	18	23	07	07	13	03	14	11	05	03	12	07	14	15	13	03	18	18	03	11	13	20	18	21	18	15
14	25	17	17	22	24	21	14	17	23	14	15	05	09	15	09	12	11	08	11	12	20	03	03	19	19	17	09	10	07	03	16	04	03	03	14	13	16	22	23	13	17	13	15	20	20	22	16	21	26	23
15	02	07	06	11	15	14	15	10	10	07	00	01	04	03	10	11	12	01	03	07	12	06	02	11	09	10	06	03	-05	01	02	00	-02	04	03	10	10	13	15	09	-01	07	02	11	14	18	14	12	10	13
16	08	15	06	11	17	18	21	11	20	12	01	08	09	05	09	17	20	03	13	06	09	06	05	11	11	23	20	07	17	13	13	05	-02	01	01	-03	10	10	19	14	-01	10	07	11	15	16	18	18	18	12
17	13	15	19	26	32	26	17	17	17	16	17	15	15	17	21	20	17	10	-02	11	09	16	16	24	27	19	15	10	07	06	14	-02	-02	03	11	03	13	12	17	13	17	20	20	21	28	31	21	19	16	22
18	12	15	12	16	13	13	11	09	12	08	06	07	07	11	01	05	02	05	03	03	03	11	11	06	03	03	07	-03	02	06	16	07	01	05	10	14	18	17	09	09	09	07	05	05	04	09	03	14	10	09
19	16	15	14	14	21	18	18	13	24	15	11	-07	08	07	10	15	14	06	13	12	04	02	10	15	13	19	09	05	16	05	15	07	10	03	10	04	18	13	22	11	07	13	15	13	14	18	17	18	22	20
20	19	20	19	19	21	26	25	30	29	17	11	-02	12	13	15	14	20	24	16	17	10	15	12	18	16	18	23	14	17	17	12	05	04	10	06	10	14	17	24	13	07	13	12	14	20	28	23	30	24	27
21	-04	-01	04	-09	-02	07	04	05	12	08	-01	-01	07	-08	01	04	07	05	10	01	-01	-02	03	-04	02	09	01	08	10	-03	-01	-05	05	-06	01	02	-07	05	12	-02	-03	-06	05	-07	00	07	03	05	11	04
22	-05	05	09	07	15	10	05	06	11	04	06	07	11	00	14	15	12	10	06	03	14	14	07	02	22	18	07	09	03	06	-03	-11	-03	03	05	-02	-07	05	06	-02	-01	12	12	07	15	12	04	07	09	07
23	-02	02	04	-02	05	08	11	-01	09	09	06	03	10	-04	02	07	05	02	07	03	14	07	04	-06	10	13	03	-05	03	06	-01	-02	-06	-02	07	-01	11	05	13	04	-06	02	07	-02	02	06	11	-01	09	08
24	03	09	02	06	10	08	-05	04	10	09	10	02	-02	-01	02	08	-04	-05	06	02	14	05	00	00	11	09	-05	10	07	10	03	-03	-03	-02	09	11	03	03	05	05	03	04	01	07	10	09	00	01	10	08
25	01	-05	07	-02	07	10	02	02	10	08	04	04	08	-02	03	12	01	01	11	07	08	03	05	01	11	13	02	-01	07	03	01	-09	-05	-05	04	01	05	06	10	04	03	01	06	04	13	11	05	05	10	10
26	01	05	00	07	06	03	08	04	06	04	-01	01	02	00	06	11	16	03	08	03	09	09	01	03	10	04	06	12	-03	00	03	-07	14	06	-03	-03	00	-03	05	-05	01	03	03	05	11	07	13	05	03	05
27	20	14	11	07	19	14	09	13	03	09	21	12	16	02	14	13	09	10	00	04	12	09	03	07	22	19	13	05	09	07	09	-02	-03	07	03	00	03	03	04	06	12	18	14	07	14	14	12	10	00	10
28	07	06	08	18	26	24	15	14	14	09	05	09	07	15	14	24	19	10	15	02	03	05	-01	10	20	22	11	02	12	12	17	-01	-02	01	10	03	06	07	09	03	09	15	11	19	26	26	19	13	15	13
29	-07	-10	-10	-13	-08	-04	05	-04	-06	-05	-02	-06	-02	-12	-03	-02	08	-09	-06	-04	-03	-01	-02	-09	-03	-03	05	01	07	06	-01	-07	00	00	-02	01	07	02	-05	-11	-04	03	-04	-18	-01	-11	03	-02	-03	-12
30	02	10	12	15	16	15	00	08	06	00	-01	00	14	16	12	17	-02	06	06	-05	00	06	05	17	18	16	00	00	-07	-12	05	02	-02	-10	06	00	-04	04	10	03	11	02	14	18	17	18	02	07	06	05
31	15	17	11	15	19	11	09	04	11	03	01	-04	10	16	19	16	15	05	04	-03	10	10	11	18	21	19	14	11	05	00	10	-01	-05	-06	-06	-11	03	01	17	06	05	16	17	18	20	14	15	11	09	09
32	-01	08	03	10	18	05	04	10	13	10	-03	-04	06	05	15	06	11	09	03	02	02	03	02	13	22	06	12	16	09	06	10	01	-05	-03	00	-06	02	03	11	03	-06	05	04	10	21	06	09	12	13	-08
33	16	11	16	09	18	12	06	09	15	18	12	-01	14	07	08	10	03	-03	11	11	01	02	10	12	15	12	03	02	03	06	14	03	00	01	03	04	10	04	17	14	10	13	19	13	16	16	10	09	14	22
34	04	06	11	16	23	08	03	02	08	01	-02	10	20	22	20	13	06	-02	06	-06	11	10	12	16	25	18	07	02	02	-11	11	-06	-13	-12	-03	-02	03	05	04	03	09	14	13	21	27	13	06	01	06	01
35	08	14	11	18	26	21	13	19	19	12	08	03	11	16	16	24	20	04	20	09	03	12	03	18	21	20	16	06	03	06	14	04	05	-01	03	03	05	16	23	15	10	13	15	21	27	26	18	19	20	18
36	18	17	28	18	25	24	25	28	31	26	12	06	20	11	22	15	22	19	16	15	10	12	23	17	25	20	26	20	25	17	11	02	-03	03	06	10	18	12	27	22	12	19	26	21	21	26	23	25	32	28
37	-03	05	19	15	20	21	21	17	28	26	06	02	07	06	12	12	06	07	14	09	09	04	15	11	16	12	10	14	17	15	03	-05	-01	01	07	21	24	16	29	28	03	12	18	14	18	23	23	20	28	29
38	00	02	-01	01	05	07	05	13	05	16	-02	-08	04	-01	09	04	06	15	09	18	03	06	01	04	05	-02	-05	04	10	10	03	-03	02	02	02	03	03	04	-04	14	-03	02	03	-05	04	04	09	14	08	15
39	08	02	11	13	20	24	15	23	28	30	08	-01	07	15	15	20	13	08	17	22	11	07	09	18	22	24	14	11	15	12	15	-04	-10	-07	-03	05	13	23	32	25	12	11	14	14	20	29	19	27	28	33
40	09	19	09	12	23	28	24	22	24	27	10	09	12	03	21	26	17	15	17	17	13	19	14	13	27	22	24	16	27	18	03	02	05	02	03	13	20	12	23	26	03	14	03	11	24	31	28	23	20	29
41	-03	-01	07	10	02	02	06	04	04	-03	-03	02	08	17	-01	08	04	04	09	-03	15	03	02	10	06	07	07	-02	01	-09	-05	-05	-06	-07	-01	-05	-01	03	-01	-02	13	-02	12	11	09	05	10	05	06	-04
42	01	00	01	01	-01	07	-02	-01	15	12	03	04	-05	06	04	08	02	-02	18	01	08	07	-02	-02	00	03	00	03	10	02	-03	-07	04	-05	-06	-06	-05	09	10	05	15	01	-04	10	05	03	03	03	15	09
43	-03	04	17	14	03	08	03	13	17	17	02	03	05	13	02	06	-03	-01	14	15	10	02	10	04	03	06	-02	02	-01	05	04	-04	02	05	17	07	05	17	21	18	02	05	03	03	04	07	04	16	16	18
44	00	-01	01	06	05	02	06	04	09	04	00	05	00	08	03	06	07	-01	09	03	17	03	-05	-01	03	10	02	03	-01	03	05	-07	06	-03	-02	-07	-03	03	03	06	05	03	07	08	10	10	04	05	05	07
45	-07	06	00	07	14	03	03	09	10	03	-02	06	09	08	11	00	00	08	07	-03	03	-01	-04	13	15	-02	03	12	04	-04	05	-01	-04	-03	03	10	07	-01	09	01	05	02	05	10	14	08	05	11	09	07
46	00	07	10	08	08	13	20	16	22	15	10	13	09	06	05	12	15	14	20	17	13	10	09	09	04	06	12	10	07	17	-07	-07	-03	-03	05	03	04	07	14	11	04	13	03	10	07	15	20	17	19	19
47	02	10	07	18	20	30	24	20	23	21	15	01	01	16	19	24	27	16	18	21	21	16	08	16	23	27	16	20	07	16	02	-02	-01	06	04	11	15	13	19	11	20	12	11	20	25	32	29	24	21	23
48	-02	06	-08	01	03	-02	-02	-07	00	-07	-01	05	01	01	04	00	01	-05	06	-07	17	10	-05	00	05	02	00	05	03	-06	-01	-08	-05	-11	-03	-01	-01	-07	-03	-03	-11	-05	-06	02	02	-05	-05	-11	-04	12
49	10	11	09	17	26	28	16	18	18	19	18	14	14	15	19	29	19	12	16	14	18	15	15	19	27	23	12	10	09	07	13	-04	-02	-05	06	01	04	08	23	12	13	16	16	18	28	30	21	20	19	25
50	03	06	04	06	03	05	06	06	03	-04	-07	03	-06	00	-07	-07	-09	-01	00	-07	04	10	01	-05	-05	04	00	-02	-04	08	08	03	-01	03	07	12	14	10	07	05	08	07	05	03	04	10	07	07	06	07

[a] See Appendix D for identification of reference measures.

APPENDIX F

INTERCORRELATIONS, MEANS AND STANDARD DEVIATIONS AMONG PERFORMANCE MEASURES

VARIABLE	No.	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	Mean	*SDs*	No.
Integrated Error Score 1	51		68	26	20	15	10	07	13	07	04	25	08	-01	-04	-04	00	05	03	04	00	15	05	02	02	02	05	08	09	04	-03	60	50	26	25	25	13	13	09	15	09	35	25	01	09	05	03	05	10	05	08	5.06	1.93	51
Integrated Error Score 2	52			43	37	29	19	18	23	13	11	24	29	12	08	11	05	10	12	11	09	21	36	22	18	18	10	15	18	08	05	36	37	11	19	32	22	19	12	16	15	37	45	18	23	20	12	13	20	11	17	5.04	1.90	52
Integrated Error Score 3	53				61	52	36	20	31	20	16	16	29	58	47	37	27	09	19	19	16	10	27	67	43	40	31	20	18	11	13	12	06	-05	03	24	17	18	17	18	18	33	42	70	52	47	34	17	29	21	18	4.99	1.98	53
Integrated Error Score 4	54					73	62	36	41	27	13	26	36	48	70	58	50	32	30	17	14	26	40	52	77	60	51	29	27	18	14	04	-04	-13	-05	15	16	19	25	25	20	32	40	53	83	66	58	34	40	26	19	5.01	1.92	54
Integrated Error Score 5	55						63	41	40	32	30	24	34	48	54	76	57	36	33	20	21	27	43	50	58	79	56	37	28	26	22	08	-08	-18	-03	11	14	21	24	29	28	29	34	55	69	86	62	40	42	34	32	5.00	1.94	55
Integrated Error Score 6	56							61	54	47	41	22	23	29	50	55	73	48	39	32	28	26	36	35	54	55	74	53	39	37	31	-03	-04	-14	-09	12	35	33	39	41	33	24	27	42	60	63	92	57	56	48	43	4.96	1.96	56
Integrated Error Score 7	57								65	60	46	-01	12	06	19	25	38	65	48	45	37	11	21	12	26	32	37	72	43	47	42	06	09	12	18	24	38	63	47	48	41	09	16	21	28	38	59	92	63	60	52	4.96	1.95	57
Integrated Error Score 8	58									71	57	03	04	10	26	23	30	38	67	59	53	08	14	12	34	32	33	53	68	57	46	17	14	15	14	25	42	47	66	57	48	11	15	19	30	35	48	63	92	70	62	5.02	1.99	58
Integrated Error Score 9	59										70	-02	-06	06	19	21	31	41	47	76	57	12	11	05	27	25	31	50	48	72	58	14	11	09	04	14	33	37	58	77	57	03	08	16	25	30	49	59	73	95	74	5.00	1.93	59
Integrated Error Score 10	60											02	-07	03	11	19	27	28	34	49	72	08	10	12	17	26	24	36	42	52	63	12	04	-04	00	13	30	34	51	65	74	-02	03	12	13	25	42	48	63	73	90	5.02	1.97	60
Horizontal Error Score 1	61												46	20	17	22	19	04	00	-05	04	31	30	26	18	29	22	00	00	-06	-04	-05	-09	-21	-13	-02	-01	-03	01	08	00	46	40	23	22	19	17	00	03	-02	06	4.99	1.99	61
Horizontal Error Score 2	62													37	31	28	30	09	06	-06	-09	25	42	34	25	37	21	13	-03	-06	-10	-15	-24	-26	-19	09	04	-01	00	02	-03	39	48	39	37	37	26	13	03	-03	05	5.03	1.89	62
Horizontal Error Score 3	63														60	56	41	17	16	08	00	21	34	68	57	54	39	19	03	03	08	-11	-30	-50	-39	-15	-14	-08	-04	04	01	25	35	71	55	53	33	07	08	07	03	5.00	1.99	63
Horizontal Error Score 4	64															68	61	28	27	12	09	21	33	56	80	65	54	24	12	10	10	-13	-25	-40	-51	-25	-15	-01	14	19	13	27	36	62	74	62	55	22	27	20	12	5.00	1.94	64
Horizontal Error Score 5	65																68	39	26	12	13	27	47	50	66	78	59	32	19	19	17	-13	-25	-38	-38	-32	-12	-05	08	19	16	24	35	55	65	81	58	27	24	21	17	4.96	2.00	65
Horizontal Error Score 6	66																	47	26	18	16	24	38	37	36	62	72	40	20	25	20	-10	-21	-29	-39	-21	-08	02	24	28	17	24	33	47	58	66	77	40	35	30	24	5.02	1.93	66
Horizontal Error Score 7	67																		53	39	26	14	23	15	36	39	48	68	36	40	35	04	-05	-05	-07	-10	-02	02	12	22	15	05	16	23	31	40	50	64	40	39	28	4.97	1.95	67
Horizontal Error Score 8	68																			56	32	01	12	15	31	27	30	48	59	55	39	04	02	02	-05	05	18	10	18	20	18	02	06	19	25	31	37	49	58	48	34	5.00	1.95	68
Horizontal Error Score 9	69																				56	-01	07	04	15	15	18	41	49	68	56	14	10	11	04	04	25	14	32	41	26	-02	04	13	17	21	32	44	55	71	50	5.00	1.93	69
Horizontal Error Score 10	70																					02	13	10	14	19	17	28	39	49	66	10	03	00	08	13	22	27	36	45	39	-04	07	07	08	20	28	39	53	58	70	5.05	1.95	70
Vertical Error Score 1	71																						48	18	19	32	30	11	00	00	02	-16	-25	-12	-08	-08	02	02	16	20	09	31	26	20	30	29	28	12	10	11	13	5.07	1.90	71
Vertical Error Score 2	72																							41	34	47	36	24	07	05	09	-23	-35	-27	-18	-07	03	01	11	15	01	32	49	40	39	47	37	21	16	11	17	4.96	1.97	72
Vertical Error Score 3	73																								55	55	35	22	11	00	08	-13	-24	-49	-31	-05	-07	05	01	08	09	31	43	72	55	51	35	12	13	07	09	4.97	1.95	73
Vertical Error Score 4	74																									64	55	28	24	18	13	-09	-20	-34	-43	-17	-10	03	20	24	18	26	35	53	77	64	54	28	35	26	18	5.01	1.93	74
Vertical Error Score 5	75																										59	34	24	17	20	-07	-23	-35	-36	-17	-05	07	18	23	23	29	37	57	64	80	57	35	33	26	26	5.07	1.88	75
Vertical Error Score 6	76																											44	27	27	18	-07	-13	-24	-29	-15	-03	07	23	23	20	24	26	44	56	60	72	37	36	30	26	5.01	1.97	76
Vertical Error Score 7	77																												50	48	41	04	01	-10	-01	00	16	27	25	35	28	05	17	30	31	37	52	68	54	50	39	5.05	1.97	77
Vertical Error Score 8	78																													45	43	12	12	03	04	08	29	22	26	27	27	03	03	13	23	22	35	42	62	46	42	4.98	1.95	78
Vertical Error Score 9	79																														61	14	13	07	04	02	19	19	30	41	30	-03	07	06	14	24	36	44	53	68	49	5.08	1.98	79
Vertical Error Score 10	80																															04	02	-06	00	-01	25	23	28	40	40	-06	09	08	10	19	30	41	51	58	63	5.09	1.99	80
Sideslip Error Score 1	81																																59	34	31	26	08	10	10	09	09	-01	00	-09	-11	-06	-08	02	11	12	10	5.01	1.95	81
Sideslip Error Score 2	82																																	42	39	32	21	19	10	07	14	-07	-08	-26	-20	-18	-13	02	12	10	08	4.96	1.97	82
Sideslip Error Score 3	83																																		54	35	29	21	13	02	01	-10	-18	-37	-29	-26	-18	07	10	04	01	5.03	1.92	83
Sideslip Error Score 4	84																																			59	41	33	14	03	01	-05	-13	-29	-27	-24	-16	09	12	03	05	5.04	1.92	84
Sideslip Error Score 5	85																																				48	43	25	16	15	06	03	-05	-08	-06	02	19	24	15	17	5.00	1.95	85
Sideslip Error Score 6	86																																					53	35	32	32	05	-01	-05	-01	03	26	30	41	33	32	4.99	1.96	86
Sideslip Error Score 7	87																																						51	42	41	12	08	08	05	10	27	58	51	37	43	4.91	1.94	87
Sideslip Error Score 8	88																																							66	53	10	10	07	17	22	36	47	74	59	57	5.01	1.94	88
Sideslip Error Score 9	89																																								68	11	12	11	23	26	42	50	65	80	72	5.00	1.93	89
Sideslip Error Score 10	90																																									02	02	10	21	21	32	43	55	61	76	5.01	1.95	90
Time-On-Target Score 1	91																																										55	38	37	31	25	13	11	06	01	3.72	1.74	91
Time-On-Target Score 2	92																																											44	42	35	27	19	17	08	07	4.60	2.03	92
Time-On-Target Score 3	93																																												63	59	49	24	19	17	13	4.72	2.19	93
Time-On-Target Score 4	94																																													74	62	31	30	26	18	4.84	2.01	94
Time-On-Target Score 5	95																																														67	40	36	33	27	4.83	2.04	95
Time-On-Target Score 6	96																																															60	52	49	45	5.02	1.97	96
Time-On-Target Score 7	97																																																70	62	56	4.99	1.93	97
Time-On-Target Score 8	98																																																	74	70	4.98	1.92	98
Time-On-Target Score 9	99																																																		78	5.00	1.94	99
Time-On-Target Score 10	100																																																			4.98	1.96	100

APPENDIX G

Unrotated Factor Loadings

		FACTOR															
VARIABLE	No.	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	h^2
Aerial Orientation	1	60	−25	04	−06	−15	−01	−04	08	10	21	−17	−05	09	05	05	55
Complex Movements	2	55	−04	03	13	14	05	02	−02	06	−09	11	09	22	−08	−03	45
Coordinate Movements	3	57	−18	18	13	05	13	02	−05	16	16	05	05	01	01	05	49
Directional Control	4	50	−03	03	21	09	13	11	−04	23	16	02	−17	−14	06	06	47
Discrimination Reaction Time	5	50	11	07	24	−13	10	17	03	13	−16	15	−01	17	−05	00	47
Following Directions	6	40	−31	21	11	−05	07	06	−06	−19	10	12	02	−04	−03	15	41
Forced Landings	7	54	−03	12	12	09	09	07	01	15	−08	−02	−09	−36	07	00	51
Formation Visualization	8	66	−37	09	−03	15	−08	−21	13	−09	−11	02	−08	08	03	−09	71
General Mechanics	9	64	−19	−13	−19	00	−28	−10	05	−10	−08	−06	04	−13	−15	13	66
Instrument Comprehension	10	70	−25	05	06	−01	00	09	03	−01	−10	−12	−08	07	09	01	61
Mechanical Comprehension	11	60	−33	06	−15	−01	−18	−18	05	04	00	−02	−03	−17	−15	07	62
Pattern Comprehension	12	65	−25	−04	05	18	−12	−27	18	06	−06	10	08	00	−10	−06	67
Planning A Course	13	51	−17	19	07	04	17	15	04	−11	03	−12	27	03	−16	05	52
Signal Interpretation	14	49	−22	08	01	09	27	22	−07	02	07	−14	16	−11	04	−12	51
Spatial Orientation	15	45	09	10	15	00	09	−13	−06	−28	−04	−14	05	06	15	−03	40
Speed of Identification	16	41	13	23	36	−09	−07	18	01	−24	−02	17	−17	−11	−03	04	54
Stick and Rudder Orientation	17	64	−35	−02	−19	00	07	−22	−01	12	08	−13	08	09	07	−02	68
Vocabulary	18	22	−35	31	07	−34	−38	01	−19	14	07	−06	−10	−06	−04	−13	63
Visual Pursuit	19	61	−05	−01	10	02	12	02	−10	−22	16	12	02	−01	−06	−15	52
Visualization of Maneuvers	20	63	−29	01	01	−12	12	−22	00	−07	−07	−02	−14	05	16	03	62
Finger Dexterity	21	25	30	−09	19	18	−19	−16	−01	−03	10	18	−01	−08	−12	−27	42
Purdue Pegboard	22	37	40	−11	35	15	−06	−18	−03	−02	07	−06	10	−02	16	−03	52
Rate of Manipulation (placing)	23	14	51	−05	37	05	−12	−18	01	07	−07	−15	01	−04	−02	04	50
Rate of Manipulation (turning)	24	28	45	−11	26	−04	−06	−13	−08	−04	−03	−27	11	01	−12	00	49
Two Plate Tapping	25	25	45	−08	16	−12	−01	−03	12	08	02	−24	−01	−01	06	06	42
Precision Steadiness (errors)	26	40	20	04	−08	08	−11	08	20	−13	23	15	05	−07	03	14	39
Direction Control	27	57	−10	−09	07	05	09	00	08	13	−07	−03	−15	18	−15	−04	47
Control Sensitivity	28	39	35	08	−11	−18	02	11	−02	−12	−31	−11	−04	−04	−15	14	51
Ten Target Aiming (errors)	29	−03	25	34	−09	−49	25	−25	04	13	03	09	06	−13	05	−12	61
Ten Target Aiming (correct)	30	18	46	20	−11	−35	17	−23	28	−03	09	12	16	02	−05	−01	63
Bimanual Matching	31	51	15	−23	−10	−01	11	15	08	−07	08	−02	−03	−06	08	06	42
Track Tracing (errors)	32	46	20	−06	−08	11	−22	15	21	−05	07	09	08	03	18	−07	44
Two-Hand Coordination	33	51	15	−20	−21	−07	09	08	−16	08	07	01	−13	−01	−08	−13	47
Choice Reaction Time	34	21	24	23	−02	−05	−05	21	19	04	−21	−07	−09	00	14	−16	34
Complex Coordinator	35	59	24	−01	−19	−09	04	16	02	−17	−10	−05	−12	02	−09	−08	54
Pursuit Confusion (errors)	36	50	−02	−33	−02	01	02	04	−16	11	−09	16	15	−04	13	04	47
Pursuit Confusion (correct)	37	48	15	−29	−10	−06	03	07	−23	08	−18	10	17	−15	−05	−04	51
Motor Judgment	38	17	12	−22	−02	−12	−21	20	−14	15	10	−04	13	05	05	−03	27
Discrimination Reaction Time	39	52	18	10	−03	14	09	−03	−03	15	−24	14	03	04	07	05	45
Rotary Pursuit	40	43	24	−34	−04	−16	−09	−05	06	16	09	11	−09	09	02	24	51
Visual Reaction Time	41	−03	28	44	−14	17	−02	17	12	21	04	01	00	15	−09	01	45
Auditory Reaction Time	42	09	26	35	−17	29	03	03	−08	11	12	−13	05	−07	−18	01	40
Jump Visual Reaction Time	43	11	38	36	−14	12	03	−18	−28	00	−04	16	00	04	07	15	49
Jump Auditory Reaction Time	44	22	33	31	−32	24	−12	−15	−23	−03	−01	−05	−08	−01	13	05	54
Steadiness Aiming	45	14	12	10	−31	07	−20	06	22	05	−12	02	08	−08	11	−11	29
Single Dimension Pursuit	46	24	14	−04	−09	−14	−01	−13	−32	−04	00	05	−08	12	−05	−08	26
Rate Control	47	45	21	05	−27	00	−02	09	−15	−17	12	−08	−01	09	12	−04	42
Rotary Aiming	48	07	41	11	14	03	−13	06	00	02	21	−01	−13	09	−07	00	30
Rudder Control	49	55	23	−25	−25	−15	−09	11	03	−03	13	02	01	02	−06	−03	55
Current Affairs	50	09	−29	33	21	−25	−41	15	−12	02	−05	04	23	08	09	06	59

Note.—Decimal points have been omitted.

APPENDIX H

ROTATED FACTOR LOADINGS

VARIABLE	No.	FACTOR I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	h^2
Aerial Orientation	1	52	05	-01	08	-12	25	05	05	-02	20	26	-11	13	-07	17	53
Complex Movements	2	16	10	-15	04	19	13	20	01	10	34	30	13	04	04	19	42
Coordinate Movements	3	36	00	-05	11	15	24	19	-01	03	18	40	06	14	04	-03	47
Directional Control	4	34	13	-16	05	08	13	27	-07	01	04	30	-11	17	20	-12	45
Discrimination Reaction Time	5	07	30	-10	15	14	20	38	-06	14	13	25	04	02	-04	17	45
Following Directions	6	27	-10	-01	02	05	23	06	04	31	09	27	06	20	-16	-08	40
Forced Landings	7	29	12	-03	09	07	11	36	08	04	13	10	-03	35	20	-15	48
Formation Visualization	8	33	-11	-03	-03	-05	18	15	15	20	61	19	-06	18	04	12	70
General Mechanics	9	29	18	-04	-12	-04	19	-11	19	11	38	07	02	47	01	21	64
Instrument Comprehension	10	47	09	-16	02	-01	24	28	08	15	33	14	00	16	-05	13	59
Mechanical Comprehension	11	32	-03	07	-06	-05	21	-02	11	04	38	17	-06	47	03	12	58
Pattern Comprehension	12	16	01	-08	05	-05	16	06	17	07	58	32	00	27	16	14	65
Planning A Course	13	31	-10	-13	25	17	11	14	13	10	12	23	22	24	-16	-16	50
Signal Interpretation	14	45	-10	-12	10	06	01	30	05	-01	03	18	25	19	01	05	47
Spatial Orientation	15	34	10	02	27	10	01	06	05	30	24	-01	10	-03	06	04	38
Speed of Identification	16	09	17	-05	15	16	20	23	08	53	06	13	-10	14	10	00	53
Stick and Rudder Orientation	17	53	-04	05	02	-11	16	03	01	-12	57	25	07	17	-03	14	77
Vocabulary	18	-18	-18	13	-04	-12	67	05	-03	08	15	-02	-06	11	13	02	60
Visual Pursuit	19	35	03	00	07	04	-01	13	-02	35	14	33	14	16	16	18	50
Visualization of Maneuvers	20	46	08	07	05	-16	16	16	-08	21	47	16	-04	13	-07	01	62
Finger Dexterity	21	-08	07	-03	04	05	-01	-02	05	18	13	17	00	05	49	16	36
Purdue Pegboard	22	14	33	-13	33	15	-04	-02	07	14	16	09	06	-10	43	-02	53
Rate of Manipulation (placing)	23	-10	35	-13	38	20	00	-05	-02	03	06	-10	-09	-06	36	00	49
Rate of Manipulation (turning)	24	08	33	-13	40	18	-02	-10	-07	05	06	-10	06	04	27	19	47
Two Plate Tapping	25	14	41	-05	35	08	-04	04	04	-04	-05	-07	-13	-02	17	12	38
Precision Steadiness (errors)	26	20	22	06	-01	13	-06	-05	34	16	-01	28	-08	14	08	08	37
Direction Control	27	24	11	-19	03	04	08	23	-10	02	34	25	-12	14	00	26	45
Control Sensitivity	28	13	38	10	13	29	00	17	01	14	06	-19	01	26	-11	24	50
Ten Target Aiming (errors)	29	-06	07	63	35	-05	04	15	-13	-09	14	03	-03	-01	-05	02	59
Ten Target Aiming (correct)	30	-06	26	50	42	07	-13	01	10	00	-04	18	-07	01	-07	18	60
Bimanual Matching	31	33	37	-17	08	-04	-13	14	05	14	04	20	01	14	09	10	42
Track Tracing (errors)	32	21	29	-03	-06	04	-01	10	42	09	10	16	00	-02	18	20	43
Two-Hand Coordination	33	37	25	07	-13	04	-07	15	-16	-02	06	14	04	18	17	30	45
Choice Reaction Time	34	06	15	07	09	12	06	39	25	02	-02	-15	-11	-07	02	15	33
Complex Coordinator	35	34	28	09	00	17	-08	23	07	21	10	00	-03	20	02	38	53
Pursuit Confusion (errors)	36	23	37	-08	-15	-08	02	11	-04	00	24	21	31	14	11	03	46
Pursuit Confusion (correct)	37	17	38	-01	-11	05	01	13	-08	-01	15	05	35	28	18	16	49
Motor Judgment	38	14	28	-11	-11	-06	15	-04	02	-11	-14	04	17	-04	13	20	27
Discrimination Reaction Time	39	14	28	07	01	29	04	29	07	-01	34	15	08	08	09	01	43
Rotary Pursuit	40	17	60	-02	-06	-04	01	00	-05	-04	12	27	-14	08	04	12	53
Visual Reaction Time	41	-08	-07	13	03	48	05	15	21	-22	-16	07	-16	-13	-02	09	46
Auditory Reaction Time	42	12	-14	11	07	51	-04	-01	12	-20	-09	01	-05	10	16	05	41
Jump Visual Reaction Time	43	03	13	36	01	52	01	-04	-02	05	07	00	04	-11	10	-14	43
Jump Auditory Reaction Time	44	26	07	31	-12	48	-01	-06	13	-02	14	-16	-04	-06	20	-01	52
Steadiness Aiming	45	02	07	18	-12	04	-02	11	40	-13	08	-10	-03	05	05	14	28
Single Dimension Pursuit	46	15	15	19	-06	10	07	-05	-26	12	11	-01	08	-03	10	20	25
Rate Control	47	47	19	15	-08	21	01	01	11	15	03	01	07	-02	07	25	44
Rotary Aiming	48	-03	-01	02	12	26	09	-09	03	08	-21	11	-21	-14	17	07	26
Rudder Control	49	34	40	08	-08	00	-08	00	09	03	03	18	03	19	08	40	54
Current Affairs	50	-02	07	-02	01	-03	67	-04	19	15	-05	-01	19	-08	-12	01	58
Integrated Error Score 1	51	20	-05	-10	04	02	13	02	-08	00	01	11	-02	16	-12	06	13
Integrated Error Score 2	52	20	01	-06	03	00	12	-01	-03	10	-02	-08	-05	08	-06	12	10
Integrated Error Score 3	53	14	02	07	09	05	14	08	-11	-06	14	06	16	07	00	09	14
Integrated Error Score 4	54	23	04	02	00	12	16	13	09	01	06	04	09	-01	-17	17	19
Integrated Error Score 5	55	30	11	-02	08	01	13	22	12	-01	06	01	04	10	-06	18	24
Integrated Error Score 6	56	25	12	04	10	10	15	07	-05	00	10	07	12	09	-08	16	19
Integrated Error Score 7	57	14	08	11	-07	04	18	13	-08	13	03	07	08	02	00	28	21
Integrated Error Score 8	58	16	10	09	-07	07	14	12	-03	00	13	12	14	00	-04	22	19
Integrated Error Score 9	59	18	03	08	-04	18	18	11	-03	03	09	10	21	10	13	22	26
Integrated Error Score 10	60	11	06	05	-11	12	18	03	-10	-09	06	03	14	17	13	29	24
Horizontal Error Score 1	61	19	05	-03	03	06	16	-05	-16	-05	04	01	05	09	-01	02	11
Horizontal Error Score 2	62	05	06	-02	-01	09	10	05	-04	-02	04	-05	-05	01	-01	-03	04
Horizontal Error Score 3	63	09	13	02	02	-04	13	21	02	-06	13	10	-06	-09	05	-01	13
Horizontal Error Score 4	64	18	05	05	-07	16	09	23	09	00	04	03	03	-10	-18	04	18
Horizontal Error Score 5	65	14	12	-01	-02	-05	05	21	11	-04	11	07	-04	-06	-01	21	16

APPENDIX H—Continued

			FACTOR														
VARIABLE		No.	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	h^2
Horizontal Error Score	6	66	21	14	02	10	16	06	04	00	04	04	13	−02	09	−11	13	16
Horizontal Error Score	7	67	21	03	05	−05	03	01	14	03	16	00	07	−01	−03	−01	19	14
Horizontal Error Score	8	68	14	03	00	−05	−02	10	00	03	04	01	10	07	−07	−03	20	10
Horizontal Error Score	9	69	13	12	04	−05	19	06	−06	19	04	−10	08	10	03	02	16	16
Horizontal Error Score	10	70	19	−01	03	−07	10	09	−11	−15	04	02	00	10	03	−01	26	17
Vertical Error Score	1	71	20	06	−10	08	29	10	02	−03	−03	−02	08	10	−15	−03	14	21
Vertical Error Score	2	72	17	07	−02	06	03	16	02	00	−04	01	05	−08	−08	00	15	10
Vertical Error Score	3	73	07	09	05	00	−04	16	09	−05	−06	09	09	09	00	01	08	06
Vertical Error Score	4	74	13	08	03	−03	06	06	24	08	−03	05	17	05	−09	−15	05	16
Vertical Error Score	5	75	18	18	−08	10	04	05	19	16	−10	10	02	−02	−03	−03	15	19
Vertical Error Score	6	76	19	09	−02	14	17	10	13	−01	07	06	10	−01	08	−04	13	17
Vertical Error Score	7	77	08	03	05	−07	−01	09	24	−01	05	06	12	−07	−05	−01	23	17
Vertical Error Score	8	78	10	00	04	−03	01	00	−01	12	02	07	13	08	02	04	11	07
Vertical Error Score	9	79	01	04	02	−02	04	13	11	−10	−04	−02	18	04	16	06	16	13
Vertical Error Score	10	80	10	05	−02	−03	05	15	−06	−19	00	03	−08	08	10	03	26	16
Sideslip Error Score	1	81	12	−01	04	00	−01	13	09	11	09	−02	02	02	18	−07	09	11
Sideslip Error Score	2	82	08	−08	−04	02	−08	16	−05	03	14	−03	01	02	08	−15	−01	09
Sideslip Error Score	3	83	03	−10	06	−03	01	−06	−19	01	10	−07	09	05	15	01	−01	10
Sideslip Error Score	4	84	05	−08	−02	−03	−05	01	−10	−03	13	01	−07	06	00	−05	11	06
Sideslip Error Score	5	85	11	−04	06	15	02	08	−20	−04	04	−01	−03	19	10	00	06	14
Sideslip Error Score	6	86	09	03	08	05	−05	18	−01	−06	−07	04	−10	27	01	03	05	15
Sideslip Error Score	7	87	05	05	06	01	−03	25	14	−04	−02	−01	00	18	11	−03	18	17
Sideslip Error Score	8	88	11	−05	08	02	15	20	17	−03	00	06	04	22	−04	01	16	19
Sideslip Error Score	9	89	12	−01	12	−03	12	07	17	−02	05	08	09	19	14	15	26	24
Sideslip Error Score	10	90	03	10	05	−11	13	16	10	−12	−05	05	−01	19	11	−01	24	20
Time-On-Target Score	1	91	22	06	00	05	15	08	05	01	−16	−01	02	13	12	−21	−03	19
Time-On-Target Score	2	92	11	08	01	03	05	06	10	−01	05	09	05	11	09	−08	08	08
Time-On-Target Score	3	93	17	12	03	−01	04	07	21	−01	−07	08	07	12	−02	00	06	13
Time-On-Target Score	4	94	19	08	−02	−01	17	−01	18	10	03	07	03	07	00	−18	09	17
Time-On-Target Score	5	95	30	14	−01	02	06	02	25	17	−04	06	−02	−05	00	−07	21	26
Time-On-Target Score	6	96	27	12	04	05	15	10	10	00	−01	17	03	13	11	−07	19	23
Time-On-Target Score	7	97	19	10	05	−04	10	13	12	−04	04	04	10	08	07	−06	23	17
Time-On-Target Score	8	98	17	02	11	−06	08	18	15	02	05	11	13	20	03	−02	22	23
Time-On-Target Score	9	99	17	03	07	−03	16	14	15	−02	−01	04	07	27	11	12	21	25
Time-On-Target Score	10	100	25	05	03	−05	19	17	−03	−19	−08	06	06	13	16	05	30	31

Note.—Decimal points have been omitted.

Vol. 74, No. 17 Whole No. 504, 1960

Psychological Monographs: General and Applied

SOME VARIABLES AFFECTING PATTERN CONCEPTION

EVA I. SHIPSTONE[1]
Harvard University

This approach to the study of concept formation is based on the assumption that there is an implicit conception of structure (patterning behavior) in adult human beings and that this can be classified into hierarchically organized categories. In order to give a more definite conception of what this means, we begin with an annotated protocol recorded for a single subject. The procedure described in this sample protocol is slightly different from that used in the main experiments described below; it was chosen, however, in order to unravel for the reader in some detail the process of strategy construction in concept formation. This technique is replaced by a simpler and more concise one in the experiments to be described later.

AN ILLUSTRATIVE STUDY
(Experiment I)

The subject sits at a table; across from him is the experimenter. The subject has just finished reading the following instructions:

> During the last war an army base was receiving important code information, one message at a time. This information had to be sorted for filing as it was received. Now suppose you were the officer on duty with this important job of receiving and filing this information. You will now be presented with the same information, single message on single card, two cards at first, then one card at a time. You are to study the card and then decide how you are to file it. You can put the cards into as many piles or boxes as you wish and you are free to move around a card from one box to another any time you wish. You will not be limited in time since the decisions you make are very important, but work rapidly. Read these instructions again until you understand them thoroughly and read them again in between if you should forget them.
>
> Any questions?
>
> Now, one further point. As you work with these cards, think aloud as if you were talking to yourself. Talk freely about what you are doing throughout your performance.
>
> Any questions?

"Here," says the experimenter, "are the first two cards." The subject is handed two cards, 5″ × 8″, on which are typewritten: ZIR PAG ZIR ZIR PAG NEN; and ZIR PAG NEN PAG NEN PAG NEN PAG NEN PAG NEN PAG NEN PAG ZIR NEN. Only three nonsense words, NEN, PAG, ZIR, arranged in various different orders, are typed on any of the cards; it is convenient to abbreviate them, N, P, and Z. The subject looks at the two cards, laughs self-consciously, and places them separately on the table in front of him.

He says, "It seems as though the main difference in the way things repeat is first according to the symbols that appear and then in the order in which they appear."

There is a silent pause. The experimenter says, "Now you will have to tell me when you are ready for the next card."

"Next card, please."

The next card contains the message ZPZN. The subject places it on the table by itself. All three cards are separately placed, so that the experimenter asks, "Why did you place that card there?"

"Oh, well," the subject says, "there seems to be a difference in the lengths of the messages. And I don't have enough information to really decide on categories. I am just setting them aside until I gather a little more information."

The experimenter, who has seen some people base their entire classification on the length of the messages, makes a special note that the first basis of classification mentioned by the subject is length. Will he stay with this aspect, or will he look further?

The next card is ZPZZZZZZZZZZZZZPN. The subject says, "Now this card is more similar in length to one of the other ones I received, and it contains a large number of ZIRs as the other one, so I will put these two together, possibly in the same category." He places together ZPZZZZZZZZZZZZZPN and ZPZZPN; since he does not combine ZPZZZZZZZZZZZZZPN and ZPNPNPNPNPNPNPZN, he seems to have rejected length as a basis of classification. For the moment, apparently, his attention is caught by another aspect, the repeating unit ZIR.

---

[1] The author, now teaching at Isabella Thoburn College, Lucknow, India, wishes to acknowledge her indebtedness to George A. Miller for directing this research. Without his invaluable help this thesis could not have been written. Gratitude is also expressed to E. B. Newman, Chairman of the department, for his suggestions.

The next card is ZPNPZN. The subject reads it silently. There is a long pause. "Whatever you are thinking," says the experimenter, "please say it aloud. I know that is not easy."

"Well," says the subject, "I just noticed that all of the five cards start with the same syllable and the syllables appear to be ordered. Similarly, the second syllables in all the cards are the same. A difference arises in the third syllable, so there is a possibility of arranging these cards according to the syllables. Alphabetically, in other words. But for the time being I will put this card in with the one which has similar length, that is, with the same number of syllables, like six or so." He proceeds to file them by length, as follows: ZPNPZN and ZPZZPN are put together, ZPNPNPNPNPNPNPZN and ZPZZZZZZZZZZZZZPN are put together, and ZPZN is left by itself. Although he is beginning to pay attention to the way the messages begin, and to consider the possibility of an alphabetical ordering, he rejects it in favor of a classification on the basis of length. He has destroyed his first tentative grouping on the basis of a repeating unit and has returned to the earlier hypothesis.

The next card is ZPPN. "Again," says the subject, "all the cards start with the same two syllables. But I am going to set this card aside for a separate category for no other reason than that I want to gather more cards before I make a classification." There is a pause. "Am I to assume that this is coded information? I mean, information like language information?"

The experimenter replies, "I am sorry I cannot answer that question for you. These are messages you are receiving and you are trying to file them."

"Messages are coming in and I am trying to organize them. Hmm . . . temporarily, I am going to rearrange the cards and put them in alphabetical order according to the syllables, starting from the beginning." The subject seems to hope that something not presently obvious will appear as he tries different spatial arrangements of the cards. As he works at alphabetizing them, he says, "I've just noticed something that had not occurred to me before, that there are only four symbols." In fact, of course, there are only three. "Somehow," he continues, "in the beginning I assumed there were going to be more. So, now I am arranging them first of all according to syllables, that is, the syllable order and the first three syllables, and then according to the lengths of the messages." Since the first two syllables are identical on all the cards that he has seen, any attempt to alphabetize must take into consideration at least the first three syllables of the message.

The experimenter notes that the subject is beginning to combine his rules in an effort to find some new basis for classification. Like most subjects, this one began with relatively simple rules—counting and alphabetizing—ready-made operations he has learned in other contexts; he transfers them bodily to the present task. For some reason, however, the results do not satisfy him and he is now beginning to search for something else. As he collects more cards, he wants to come to terms with them in a way that respects their peculiarities and particularities, not by some ready-made rule that might fit almost anything.

As shown in the chart below, the next card is ZPNPNPNPNPZN. The subject studies this card in silence for almost 60 seconds, looking back and forth between it and other cards spread out on the table before him. Then he makes a new observation, "I notice now that when I look at these cards in categories that there are units within each card that seem to repeat between cards. Like, for instance, I have two cards here both of which start with ZIR PAG NEN and end in ZIR NEN, and in between they have the unit PAG NEN, but in slightly different numbers. I mean, one card will have more of these units than another card. It seems like relationships are starting to emerge."

The more complex rule that the subject has just discovered is one that the experimenter has watched many subjects come to; the experimenter calls such formulations "Initial-Middle-Final Rule." Although the rule is here applied to only two cards to form a single category, it contains the germ of the filing system that this subject will use for all the remaining cards. Note that it is a rule designed to fit the present situation and to use the particular information provided by this set of messages; it is not a rule that would apply equally well to all imaginable sets of messages. The order in which the messages were presented to the subject was:

1. Z P Z Z P N
2. Z P N P N P N P N P N P N P Z N
3. Z P Z N
4. Z P Z Z Z Z Z Z Z Z Z Z Z Z Z P N
5. Z P N P Z N
6. Z P P N
7. Z P N P N P N P N P Z N
8. Z P Z Z Z Z Z Z P N
9. Z P Z Z Z Z Z Z Z Z Z Z Z Z Z Z P N
10. Z P P P P N Z N N N N N P
11. Z P P P N Z P
12. Z P N Z N P
13. Z P N P N P N P Z N
14. Z P P N Z P
15. Z P Z Z Z Z Z Z Z Z Z Z N P
16. Z N Z N N N N P
17. Z P N P N P N P N P N P N P N P N P Z N
18. Z P Z Z Z P N

19. Z P Z Z Z Z Z P N
20. Z P N P N P N P N P N P N P N P Z N
21. Z P P P P P P P P P P N Z N N N P
22. Z P Z Z Z Z Z Z Z P N
23. Z P Z Z Z Z Z Z Z Z Z P N
24. Z P N Z N N P
25. Z P N P N P N P N P N P Z N
26. Z P Z N
27. Z P N P N P N P N P N P N P N P N P N P Z N
28. Z N Z P
29. Z N Z N P
30. Z P P P P P N Z P

At this stage the subject has received seven cards and has arranged them into three sets as follows, which we shall label as A, B, C, for convenience:

A. ZPZZPN
ZPZZZZZZZZZZZZZPN
B. ZPNPNPNPNPZN
ZPNPNPNPNPNPNPZN
C. ZPPN
ZPZN

The seventh card, ZPNPZN, is left isolated for the time being.

The next card is ZPZZZZZPN. "Again I see the same thing within cards," the subject says, apparently with some satisfaction. "I see units repeating in the cards within each class. For example, there is an end unit, two end units, in fact, and the intermediate units are the same, just varying in number. In other words, I count the number of ZIRs in the middle unit in both cards. It may be five in one case and six in another, and this holds for all my cards, I think. In other words, in all cases the ends and the beginnings are the same. It is the middle that varies."

The next card is ZPZZZZZZZZZZZZZZPN. The subject says, "This classification seems to be working out very nicely, so I will continue."

The next card is ZPPPPNZNNNNNP. There is a long silent pause; clearly, something new has appeared. The experimenter reminds the subject to keep talking.

"All right," he says, "now we have another card. I am trying to see what category this will fit in. I am looking at the first three syllables and again I see the same pattern emerging, I think. I notice now, as a side observation, we seem to have only three basic categories. And each of these categories is divided into subclasses which consist of cards having the same beginnings and endings and the same intermediate units, varying in number." The subject places this card, ZPPPPNZNNNNNP, together with ZPPN and ZPZN.

The next card is ZPPPNZP. "Again, the same pattern. I mean to collect more cards and see if it continues. And as soon as I get a few more cards to form my third category I am going to see if there are any more relationships observable between the major categories themselves."

The next card is ZPNZNP. "Oh," says the subject, "now I notice that I made a mistake. I see that whereas I thought there were three categories I now see there are four. The end units of two of my cards in one of my classes is different." He indicates Set C consisting of ZPPN, ZPZN, ZPPPNZP, and ZPPPPNZNNNNNP. "In fact, there are considerably more categories than I thought. My first category (Set A) is all right. The beginnings and ends are the same, and the intermediates are the same. Similarly, my second category (Set B) is all right. Just this one I was trying to make my third category seems to be wrong. It does not seem to follow the same pattern the other two do. I just assumed it did, but I see that it doesn't. I could attempt to follow the same classification, that is, instead of having three categories, I could have five or six. In other words, I will continue to see if a pattern emerges. It looks like I will continue with this new category. It appears, though, that I may have to discard my old standards, or criteria."

The subject has been given no hint as to how many categories there should be; concern with the number is his own attempt to define the terms of the task. The desire to have as few categories as the evidence will allow seems to be an important consideration governing many subjects' satisfaction with their own efforts. Moreover, like most subjects, he wants to use the same rules for every category; it feels inconsistent, or at least untidy, to have a different description for each different set of messages.

Note here, incidentally, that it is his third category with a double variation in the middle that is causing his troubles.

At this stage the subject has received 12 cards and has arranged them as follows (though not necessarily in the same order within classes):

A. ZPZZPN
ZPZZZZZPN
ZPZZZZZZZZZZZZZPN
ZPZZZZZZZZZZZZZZPN
B. ZPNPNPNPNPZN
ZPNPNPNPNPNPNPZN
C. ZPPN
ZPPPNZP
D. ZPNZNP
ZPPPPNZNNNNNP

Two cards, ZPZN and ZPNPZN, are left as odd cards, not yet assigned to any definite category.

The next card is ZPNPNPNPZN. The subject says, "I just noticed something I hadn't noticed before: that all the cards begin with ZIR." Has he rediscovered something he remarked on first after the fifth card? Or is he noting that his earlier generalization is still valid? He says, "There no longer seems to be the tendency to have the same end units." Perhaps he means that the beginnings of the messages are still predictable, even though the endings are beginning to seem more variable. Unfortunately, it is not always possible to understand what a subject means; to interrupt him by questions would interfere too much with the processes the experimenter is trying to observe.

The next card is ZPPNZP. "I noticed that the card I just received is like the first card in one of my categories, the simplest card, only this has additional symbols at the end." He places ZPPNZP together with ZPPN and ZPPPNZP. "This could not occur with any of my other cards," he says. "I will continue." He means that the endings of two of his cards in a category could be different.

The next card is ZPZZZZZZZZZZNP. "Now this card fits into one of my earlier categories. I will put it there."

The next card is ZNZNNNNP. There is a long pause. Then, "Now I am lost."

"Why are you lost?"

"I don't see relationships occurring between these last five or six or seven or eight cards. I cannot see where I can fit them sensibly into the categories. They don't seem to be fitting into the same nice pattern that the earlier cards did." He leaves this card unassigned to any of the groups.

The next card is ZPNPNPNPNPNPNPN-PNPZN. "This one fits into one of my earlier categories. What I would like to see . . . I would like to get more information in order to decide whether or not I have to abandon my older classification. Maybe these cards that I have will fit better if a card could be used to establish a new category using the same criteria as before. It is not apparent, however, whether this will be useful in sorting these materials. We have sort of new materials here. The two categories that I have seem to be filling in very nicely; they seem to be complete series, whereas these other cards don't fit anywhere, yet."

The cards that are giving our subject so much trouble all belong together; the confusion is arising because there are two variables in this set as opposed to only one variable in each of Sets A and B.

While the subject was talking he reached for another card, ZPZZZPN, which he placed in one of his earlier categories.

The next card is ZPZZZZZPN. He greets it like an old friend and disposes of it quickly. "Now this one, too, fits into my older category, so I will place it there."

The next card is ZPNPNPNPNPNPNPNPZN, another old friend. "And, similarly, this one fits in all right."

The next card is ZPPPPPPPPPPNZNNNP. A stranger, alas. "Now, I have this card. . . . I see another pattern emerging, I think. . . . There are . . . now, let me think . . . yes, it's just that I didn't discover what the intermediate unit was in some of these, assuming it to be two syllables, like the more obvious one in the other set," he says, comparing this card with the one just preceding it. "I see now that the end syllable consists of just one—I mean, the ending consists of just one syllable, and the intermediate ones are more than one syllable. I can form another classification, using the beginning three symbols and ending in one symbol, at least for some of my cards. I can fit them into my classification. In other words," he says, revealing once more his desire to find a consistent basis for the classification, "I have formed a third category which is similar to the other two larger categories I have (meaning Sets A and B). And I still have four cards left over." He refers to ZPPN, ZPZN, ZPNPZN, and ZNZNNNNP as the four that are left over; then, returns to his new, or third, category in search of a simple rule: "Now, let's see if we can see the same sort of relationship here. Ah, it seems to be emerging. There is an end unit which is constant and there seems to be a middle unit which seems to be constant. So we can have two ends and a middle; then, in between ends and the middle, there are other units which are the same. The same unit is added, only just in different numbers. So we have the same sort of process as before, except a little more complex. Now the question is, how to classify these cards. We will make a category out of all the cards that have the same two end units and the middle unit and I will classify them in order of their relationships depending upon the number of units in the beginning, and then the number between the end and the middle. If there is one unit, it will be the first card; if there are three units, it will be the second card, assuming that there isn't one with two units in between. And if I have two cards with the same number of units repeating in between the beginning and the middle units, then I will use the second side—I mean, the number of syllables that is between the middle and the end is my second way of classification. I will rearrange these now."

Notice how the subject distinguishes between variant and invariant units in the statement of his rules. The approach might now be characterized as Initial-x-Middle-y-Final, where x and y represent the variable parts. And he also has the rule well enough in mind to use it for ordering the cards within each category.

"When I come to rearrange these," the subject says, "I see this isn't the case."

"What isn't the case?"

"That the two ends and a middle unit are constant with all the cards, and there is an extra syllable. So, I am going to wait."

At this stage the subject has received 21 cards and has rearranged them as follows (though not necessarily in the same order within classes):

A. ZPZZPN
ZPZZZPN
ZPZZZZZPN
ZPZZZZZZPN
ZPZZZZZZZZZZPN
ZPZZZZZZZZZZZZZPN
ZPZZZZZZZZZZZZZZPN

B. ZPNPNPNPZN
ZPNPNPNPNPZN
ZPNPNPNPNPNPNPZN
ZPNPNPNPNPNPNPNPZN
ZPNPNPNPNPNPNPNPNPNPZN

C. ZPNZNP
ZPPNZP
ZPPNNPP

ZPPPPNZNNNNNP
ZPPPPPPPPPPNZNNNP

Four cards, ZPPN, ZPZN, ZPNPZN, and ZNZNNNNP, remain unclassified. Although he is making progress, he does not feel completely confident about the third category. [At this point the reader should try, if he has not already, to fit the four odd cards into the existing categories; he will discover that this subject, like the majority of subjects, finds it difficult to invent zero when he comes to count the number of times a unit recurs.]

The next card is ZPZZZZZZZPN. "This one fits into one of my older classifications, and in my first big category."

The next card is ZPZZZZZZZZZZPN. "And this one, too. I will put it there."

The next card is ZPNZNNP. "Now I notice these . . . here . . . no, I'll see if it fits into one of my categories and if not, I will put it with these other cards and see if I can establish a new category with them. Now, this last card seems to be similar to two of my cards that I had set aside." He puts ZPNZNNP together with ZPPPPNZNNNNNP and ZPPPPPPPPPPNZNNNP. "They have the same beginning and the same end and pretty much the same middle unit, but not quite, however."

The next card is ZPNPNPNPNPNPZN. "This fits into one of my older categories."

The next card is ZPZN: it is the same as the third card he had received earlier [pause]. "This one I will place with that and that," he indicates [a ZPNZNP and ZPNZNNP. "There's . . . [a pause] it just occurred to me that, although there appears to be a relationship, it is probably only by virtue of the fact that there are only four symbols we are using." The mistake persists; he still does not realize that, in fact, only three symbols are used. "However," he continues, "there does appear that there is an order. I mean, if I were to divide it like I have two end units and middle units, it appears that there is another syllable appearing in between the beginning and the middle and it is always the same unit. It just varies in number, much like the other classification. So now I will order those. As a matter of fact, I see that three of my cards fit very nicely. actually, there are four that fit in very nicely. Five . . . oh, no, this is a little different. The end unit is different, so I have now established the fourth category. And I still have two cards left over that I am not quite sure what to do with. Let me check now and see if they don't fit somewhere . . . no, they don't seem to fit."

The next card is ZPNPNPNPNPNPNPNPNPNPZN. "This fits in very nicely with one of my earlier categories." He disposes of it quickly.

The next card is ZNZP. "This card does not seem to fit anywhere."

The next card is ZNZNP. He pairs it with ZNZNNNNP, saying, "Now I have two odd cards that seem to fit to form another category, using the same idea of picking out syllables interspaced in between, just repeating. There is just one thing I want to say pertaining to one unit. . . . So, I now have five categories and two extra cards which, temporarily, I will call separate categories."

The next card is ZPPPPPNZP. The experimenter says, "This is your last card."

The subject says, "This one, I think, fits into one of my earlier categories. Now I want to look over these couple of categories and see how I stand on those. I want to make sure that I have really followed the rules I think I established."

"You can take all the time you want to," says the experimenter, "but don't forget to talk."

At this stage the subject has received 30 cards and has arranged them as follows (though not necessarily in this order within classes):

A. ZPPN
ZPZZPN
ZPZZZPN
ZPZZZZZPN
ZPZZZZZZPN
ZPZZZZZZZPN
ZPZZZZZZZZZPN
ZPZZZZZZZZZZPN
ZPZZZZZZZZZZZZZPN
ZPZZZZZZZZZZZZZZPN

B. ZPZN
ZPNPZN
ZPNPNPNPZN
ZPNPNPNPNPZN
ZPNPNPNPNPNPZN
ZPNPNPNPNPNPNPZN
ZPNPNPNPNPNPNPNPZN
ZPNPNPNPNPNPNPNPNPZN
ZPNPNPNPNPNPNPNPNPNPZN

C. ZPNZNP
ZPNZNNP
ZPPPPNZNNNNNP
ZPPPPPPPPPPNZNNNP

D. ZPPNZP
ZPPPNZP
ZPPPPPZP

E. ZNZNP
ZNZNNNNP

Two cards, ZNZP and ZPZN, remain unclassified.

"Okay," the subject says. "Well, I'll start out with my biggest and main categories because the relationships are clear, so I'll make sure that they follow all the way through. Now, they all have the same beginning and end syllables. There is no question about that. And, similarly, the second and next-to-last symbol in all—the beginning and end units are similar in all cards. In my first category the intermediate symbols are always ZIRs and they occur in varying numbers. Let's see, one, two, three, skip some; I can order these according to the number of ZIRs in between. I don't suppose it matters as long as I am classifying rather than trying to interpret the information. I can't seem to decide between these: they seem to be homogeneous."

He shifts to Set B. "I'll take the next one. Again, we have two syllables that are the same—I mean, they are the same from card to card, beginnings and ends. And we have in between them, rather than one syllable repeating, a pair of syllables repeating . . . and always ending in the same one. So I order these according to the number of insertions."

He shifts to the Set C. "Now I am going to look at one of the more complicated categories. Here it seems there's a beginning unit that consists of two syllables on each of the two cards and the end unit which consists of only one syllable" [short pause]. "In addition, there is a middle unit that consists of NEN ZIR, or whatever it is. And in between there are varying numbers of syllables between the beginning units and the middle unit. In this category there are varying numbers of syllables PAG, so I am going to arrange them in this order. This will be my primary ordering process. And, should there be duplication in the number of PAGs, then I would check the number of symbols on the other side of the middle part, and order them according to their number" [pause]. "I see that this does occur . . . so, they are ordered first with respect to the symbols between the beginning and the middle, and then with respect to the symbols between the middle and the end. Should there be any ambiguity in the first order . . . it appears like a very nice. . . ." He combines his next three categories into one on the basis of the rule just stated; the result is a third category:

C. ZNZNP
ZNZNNNNP
ZPNZNP
ZPNZNNP
ZPPNZP
ZPPPNZP
ZPPPPNZNNNNNP
ZPPPPPNZP
ZPPPPPPPPPPNZNNNP

"Now," he continues, "I have the cards in three main categories, and I'll see if they will hold up. Again, they begin with, and end with, the same unit; and the same pattern of middle syllables repeats."

He shifts to the remaining odd cards, ZNZP and ZPZN. "Now let's see if my last category. . . . I'm not sure what to do with that—make sure if it doesn't fit into one of the old ones . . . it doesn't as such . . ." [long pause]. I'm glancing over all the categories now to make sure that this last category doesn't fit into a subset of those" [pauses while looking]. "I think. . . ." He picks up ZPZN. "I now notice that one of these cards does fit in to the, ah, my earlier category. . . . It's curious, I didn't notice it before, but it seems to be identical with the first one in one of my categories. Well, I'll place it there."

There is a long pause while he considers the remaining odd card, ZNZP. "It seems a shame to have a one-card category. I'll look over once more to see if we can find a place to put it . . . this odd card."

After a little more search he gives up. "I will leave it as it is," he says, "I am happy with my classification."

The above technique was abandoned in all the studies to be described below in favor of a simpler and quicker one: all the cards were simultaneously presented to the subject and no running protocol was recorded as the subject sorted the strings Comments the subject wished to make on his method of sorting were recorded only at the end of the experimental session.

THE PROBLEM

This account of a subject's responses to a problem solving situation has been reconstructed from a tape recording. The subject is confronted with the problem of processing information which he received a little at a time. We can think of the subject as starting the task in total darkness, for he has almost no idea what kind of information is encoded in the messages. Whatever he does with the messages is entirely a matter of his own decisions.

After watching a dozen or more subjects perform in a similar manner, one is impressed with several things. The first is the similarity between this laboratory episode and ordinary behavior shown in dealing with our daily environment. All day long stimuli impinge upon an organism in a haphazard array; unless he can rapidly sort and file information in the right compartments his life will quickly revert to chaos—to the "buzzing, blooming confusion" of the newborn child. Man orders his environment according to rules that he develops for himself or which he learns from others; by this recoding technique he can economically handle the multiplicity and the variability of his environment for present and future use.

A second impression one gathers is that men have learned to behave in this orderly manner without any specific instructions and without any knowledge of results provided by a teacher or an experimenter. Man seems to respect structure in the world wherever he finds it, regardless of consequences. It does not seem necessary here to invoke any hedonic contingencies—unless perhaps orderliness carries its own rewards: a satisfaction with the goodness of fit. The subject is uneasy when a string of symbols does not fit one of the categories he has evolved. What one means by the fit here is more subtle, and the lack of fit less obvious, than in a mechanical or a picture puzzle; without any outside guidance or control the subject moves from a search for isolated perceptual relations into a search for the conceptual pattern underlying the stimuli that he receives.

The behavior of our subject emboldens us to argue for an implicit conception of structure in man. The argument is, of course, not new to psychology. It was first brought into lively focus by the Gestalt psychologists as part of their protest against the Wundtian tradition of elementism and against the behavioristic denunciation of everything mentalistic. The Gestalt psychologists did not share with Behaviorism its contempt for mentalistic concepts, nor its distaste for introspection. On the contrary, they were against any oversimplified mechanistic view of man; they urged the importance of value in a world of facts (Köhler, 1938). Meaningful relations (Köhler, 1929; Wertheimer, 1945), organization (Köhler, 1929), fitness of parts into an organic whole (Koffka, 1935), requiredness (Köhler, 1938)—these were some of the values they wished to restore to a world hacked up into elements.

The emphasis on relations is not a novel thought, either. Even Thomas Brown (cited in Woodworth, 1948) early in the nineteenth century, and von Ehrenfels (1937) later in that century, had stressed the importance of relations. But the notable contribution of the Gestalt school was to demonstrate the crucial role of structure, or pattern, both in the stimulus configuration and in the responses of the observer. According to them the perceptual process is essentially one of organization. The brain is a dynamic neural center where perceptual objects combine or separate on the basis of similarity, contiguity, symmetry, common fate. Köhler's conclusion (1947) that an "organism responds to the pattern of stimuli to which it is exposed" (p. 103) seems to fit the illustrative account of our subject's responses to a decision making task. We cannot help but notice how often the subject is concerned with the "fit" of the messages into some category he had established: he is continually seeking "relationships," he talks about "pattern," he is trying to fit the parts into a total scheme he has set up to solve his problem. In fact, Wertheimer (1945) might have been describing our own subject instead of the children who solved geometry problems, when he wrote: "In all these cases we have structural changes, tendencies toward the better structure, toward fitting, with the disappearance of disturbances" (p. 57).

Since the notion of structure, or pattern, as proposed by the Gestalt psychologists, seems essential for the description of our subject's behavior, it becomes the starting point of our inquiry. The notion of "pattern," however, as we use it here, is something of a special case of the Gestalt notion. We are interested in what Miller and Chomsky (1958) have called "pattern conception." For the purpose of the present experiments we define pattern as the invariance of a sequence of symbols in the context of lawful change, and we simplify its meaning by defining pattern in terms of internalized rules, or relations, a subject employs to organize novel information. Rules usually imply a notion of abstraction and generalization; they go beyond the information given. They cannot be said to be covered either by the transposition experiments such as those of Köhler (cited in Osgood, 1956, p. 495), nor by the studies of stimulus generalization conducted by the

Behaviorists (Hovland, 1937a, 1937b; Razran, 1940, 1949). In both cases the results could be explained on the basis of sensory relations alone. By the use of rules the subject searches for pattern in order to recode the sensory input into manipulable, meaningful units. Perhaps a limitation on the information capacity of the human channel makes such behavior (recoding) necessary. In any case, it is our wish to find out something about these rules or patterns by which the subject handles information. We would like, therefore, to make a systematic study of some of the experimental variables that affect the nature of the rules that subjects employ in order to search for pattern in the stimulus input.

One could begin a study of rules by asking the usual questions of What? How? and Why? Hull (1920) is one among those researchers who tackled the question What? "What does a child learn," he asked, "when he calls a certain object dog?" This question is part of a broader question: What does any individual learn when he is said to have learned a concept? To find an answer Hull required adult human subjects to learn some Chinese sounds paired with certain characters in the Chinese alphabet. Hull's conclusion was that *what* an individual learns is a rule about the common characteristics of a class of objects. He placed the emphasis on the common characteristics of the stimuli; the rule, Hull equated with discrimination learning. It is, therefore, not clear to the reader whether Hull's subjects learn a rule or whether they merely learn a name for the rule.

A second answer to the question *What?* is contained in the word relations. According to a second view what the subject really learns is some kind of a rule of relations by which he groups instances as members of a class. A search for common relations among instances might take the subject beyond the common perceptual elements to something more. For the Gestalt thinkers this something more is the "organic whole" which is more than the sum of its parts. Goldstein and Scheerer (1941) describe it as "abstraction." According to them the process of abstration is different from that of aggregation of common elements under one label. Smoke (1932) discusses this same idea as symbolic reaction. Our own approach to rules is in terms of relations and we view rules of relations as pattern conception.

The question—how rules are learned or made—has been ingeniously handled by Bruner, Goodnow, and Austin (1956), and more recently by David Reiss (1958). Bruner et al. systematically investigated the methods by which subjects categorize given events. The categories into which events are to be placed are predetermined by the experimenter's rules. It is the task of the subject to discover the experimenter's rules for placing the given events into their proper boxes. In order to categorize successfully the subject uses "strategies," or plans, by which he searches for information which would be useful to him in constructing his hypotheses about how the events are distributed into the experimenter's boxes. This stepwise process—of strategy construction, searching for information, hypothesis building, testing hypotheses—the authors describe as the process of concept attainment. Of these steps, the one of strategy construction is the most critical. Bruner et al. (1956) describe strategy as "regularities in decision-making," and define it as "a pattern of decisions in the acquisition, retention, and utilization of information that serves certain objectives (p. 54). A "host of factors" influence the construction of strategies in concept attainment, according to these researchers.

In his discussion of the construction of strategies by subjects, Reiss takes Bruner's "host of factors" for granted, and he proceeds to study how a subject constructs his "model" of what the rules that govern the stimuli might be, but Reiss studies this without constraining the subject strictly by the experimenter's rules. He varies the quantity of evidence given to the subject initially, and then observes how this amount of information affects the subject's behavior in model attainment. Reiss finds that the subject's model is in fact seriously influenced by the amount of information initially presented to the subject.

In all the studies discussed above, except the last one, the experiment concludes when the subject reaches the solution as the experimeter stipulates it. Hull's subjects continue to learn the list until the criterion is reached. The criterion, of course, means the one that the experimenter has set up. Hanfmann and Kasanin (1937) turn over a block or two to suggest to their subject the "correct" grouping which he must discover. Bruner, Goodnow, and Austin confirm and infirm hypotheses. In none of these studies is the subject ultimately free to set up a criterion independently of the experimenter. Obviously then, the question arises whether the subject will proceed to work by rules even when he is free from the experimenter's criterional constraints.

Thorndike (1931) asked subjects to draw lines equal to a specified length without giving them any information about the accuracy of their results. Yet one notes in his results that the performance of subjects began to stabilize around some standard implicit in the subjects' behavior. In other words, the subject seems to be settling on some rule which he bases upon his own performance. It is immaterial whether he does this consciously or unconsciously.

One notices a similar reliance upon a subjective standard in Reiss' subjects as they construct models. The "random" group, whose hypotheses were confirmed or infirmed according to a prearranged random schedule, did not perform randomly at all. Instead, subjects constructed models quite early and began to test these on the basis of their own hypotheses, disregarding the information they received from the experimenter.

Systematic responses, which may be abruptly shifted to new modes of systematic behavior have been observed by Krechevsky (1932), Witkin (1940), and Spence (1939) in animal subjects. By means of a four-point discrimination apparatus, Krechevsky presented an "insoluble" problem to his rats. These subjects did not run around the apparatus in random fashion, however; they developed systematic modes of responses. Later, Witkin, substantiated these results on animals run in nonspecific problem situations, but his explanation was different. Spence observed the same phenomenon in chimpanzees when he presented them with an insoluble problem in a multiple-choice apparatus.

We have mentioned these studies for three reasons: first, because all of them employ a notion of a criterion of success which is externally determined; second, that this notion of an externally imposed criterion introduces some contingencies of reinforcement for the subject; and third, that when the contingencies of reinforcement occur haphazardly, the subject disregards these and begins to develop some stable modes of responses on the basis of his own performance.

All of these observations influence the design of our research; we deliberately take our point of departure from the first two and confine our interest to the third. We stated earlier that our interest is in rules. What kinds of rules does a subject make? How are his rules different from the experimenter's rules? Do these rules depend upon the nature of the stimuli, upon a knowledge of results, or upon the subject himself, or on a combination of these factors?

These and a host of other questions clamor for recognition and investigation. Twentieth century psychology has not been indifferent to this need. We have already seen how some of the earlier studies, characterized as dealing with concept formation, were investigating rules of induction. Within the same framework, but one step further, Heidbreder (1946) introduced the idea of hierarchy in rules. The notion of hierarchy of rules implies the notion of complexity of strategies that subjects construct. In later studies the researchers were interested in the effect of rules upon learning and recall. For example, Miller and Selfridge (1950) tested their subjects on memory of passages which were graded according to their statistical approximations to English from zero to high order. The better recall of meaningful materials, the authors explained on the basis of the subject's familiarity with the rules that governed the sequential properties of the meaningful materials, not the fact of meaning itself.

Miller deals with the facilitatiing effect of rules of formation upon learning and recall in a later study (1958) in which he compares the free recall of redundant and random strings. Again he finds that even when the subject is unaware of the rules of formation that govern the structure of redundant strings he still does better in memorizing the redundant strings. Prior to this study of Miller's, some experiments had been done by Aborn and Rubenstein (1952; Rubenstein & Aborn, 1954) to study the effect of rules of contextual constraints upon learning sequences of nonsense syllables. Test passages were developed by progressively limiting the context in which the nonsense syllables could appear. Thus the rules ranged from purely random to highly redundant. These researchers found that when subjects were made familiar with the rules of formation, redundant sequences were somewhat easier to learn, but the amount of information contained in the recalled materials was not constant from passage to passage.

So far the rules we have referred to have been simple rules; they pertain to two kinds of stimuli—unit stimuli and sequential stimuli—both of which have been generated with some kinds of rules of formation. Language is a good example of the second type of stimuli. Miller's rules, as well as those of Aborn and Rubenstein, and of Reiss, deal with rules of this type, that is, with rules of formation that govern sequential dependencies.

Rules that govern sequential properties of a language also have important effects upon its statistical properties (Miller, 1951). Shannon (1949) analyzes the sequential properties of language by probability rules, a technique that provides access to information measurements as well as to organization. His rules deal with the selection of a message from a determinable number of alternative messages that could occur, and with the probability of its occurrence. His rules deal with the amount of information contained in a message. Since Shannon's concern is primarily with the communication of a message and not with its use, value, storage, meaning, history, or anything else, his rules are not of direct interest to us in the present context. His terms, however, and his techniques for measuring the amount of information transmitted through a noisy channel provide the psychologist with some useful tools of analysis which information theorists (Attneave, 1959; Hovland, 1952; Luce, 1954; Miller, 1953; Quastler, 1955) have applied for measuring organization. We, too, shall draw upon these aids.

In contrast to Shannon, the communication engineer, we can place Chomsky (1957), who looks at language as a grammarian. His rules are more general and more formal than those of Aborn and Rubenstein, though not more difficult; he refers to his rules as "finite state grammars." These simpler grammars enable us to avoid the complications of natural languages and to work with simpler structures. But although his notion of rules is formally simple, it raises a few complex questions regarding relation between rules and performance. Miller (1958) points to this problem and he would "like to know what changes, if any, in a set of rules leave performance invariant," (p. 485). And so would we.

Simple rules of formation of sequential stimuli do not always produce simple effects upon responses. Some rules facilitate performance (Anderson & Leonard, 1958; Fitts, Weinstein, Rappaport, Anderson, & Leonard, 1956; Miller, 1951) and some rules hinder it (Anderson & Leonard, 1958). It seems necessary, therefore, to search for parameters that affect the construction and use of rules of formation. It is the purpose of this study to explore some specific experimental variables implicit in these rules that influence the patterning responses of subjects.

The role of specific stimulus features in *pattern perception* has been previously studied by the Gestalt group, by Attneave (1954, 1955), by Fitts and his colleagues (1956), and by others (French, 1953, 1954b, 1954c). Gestalt psychologists analyzed the peripheral properties in visual organization as proximity, similarity, continuity, and closed figure, but the process of organization itself could not, according to them, be understood without the field concept. They postulate an isomorphism between visual organization and the brain processes. According to the principle of isomorphism the environmental field tends to become "psychologically organized" in terms of regularity, symmetry, and simplicity. Attneave translated and quantified the Gestalt factors of organization in terms of information measure of redundancy in the visual input. He measured memory for simple dot patterns on recall. He found that good form, redundancy, and noise were some of the experimentally accessible variables that affect pattern perception. Fitts et al. at Ohio State University have done special work with redundant forms in visual pattern recognition. They developed metric figures of random and constrained type and they measured the amount of redundancy contained in these figures. They studied the correlation of a measured amount of redundancy with recognition of visual forms. Some additional variables, such as language habits and context (Miller, 1951), also have been found to affect the probability of response. Carmichael, Hogan, and Walter (1932), for instance, demonstrated the influence of instruction upon response. Although we shall give a different interpretation of their results later, we shall not ignore the importance of instruction as a possible source of variation in performance.

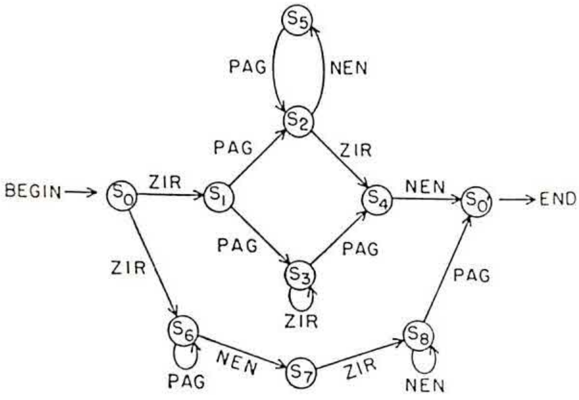
All the studies mentioned above suggest that a quantitative analysis of the experimental variables in pattern conception is possible. These studies also help to separate some parameters of pattern conception, particularly those concerned with the number of different alternatives used and with structural similarity among the sequences. The number of different alternatives has proved to be an interesting parameter in studies in perception (Attneave, 1954) and in memory (Brogden & Schmidt, 1954a, 1954b). Whether it is an important variable in pattern conception as well is the question we raise. Does a larger vocabulary help with the better ordering of the sequences of symbols or not? If not, why not? One might expect the facilitatory effect of the larger number of different alternatives to reach an asymptotic value at some point.

Structural similarity is another parameter that must be explored. Its importance in visual pattern recognition (French, 1954a, 1954b), and in learning and recall (Aborn & Rubenstein, 1952; Miller, 1958; Rubenstein & Aborn, 1954) has been demonstrated. Does it also affect pattern conception? The notion of structural similarity as used in this study should not be confused with Shannon's notion of redundancy. The languages we are working with are, in a sense, "degenerate" because in the limit their redundancy is 100%. We define structural similarity as similarity in rules of formation and specify it in terms of the number of basic strings as well as the number and placement of recursive loops within each string.

Materials and Procedure

Miller and Chomsky (1958) have pointed up some of the difficulties of method and material in laboratory studies of concept formation. In the earlier concept formation studies the subject was constrained by the nature of the task to find the rules as the experimenter defined them. Furthermore, the use of geometrical patterns differing in size, shape, and color effected a certain stereotypy in material and usually limited them to a finite number of possible alternatives from which the positive instances were to be selected. Miller and Chomsky proposed the use of a new kind of material in the study of pattern conception, viewing it as a different kind of task: instead of spatial patterns, they proposed to study sequential patterns. By using finite state systems for materials one could study pattern conception in a situation "modelled after the task of a grammarian."

A finite state language is an artificial language consisting of the set of well-formed strings (sentences) that can be generated by a set of simple rules (see Chomsky & Miller, 1958). This set of rules is a kind of algebra, or grammar, for identifying the grammatical sentences in a particular language. A finite number of symbols makes up the vocabulary of a language. A finite state grammar consists of a finite number of states (S_0, S_1, S_2, . . .) and rules for moving from one state to another. A state is defined as the set of all rules applicable to the state. Starting with any given state 0 there is a transition to a next state *i* via a symbol. Thus, for example, $S_0 \xrightarrow[a]{} S_i$ which can be written (0, a, *i*); this is one rule of transition of the grammar. A rule of a grammar determines whether there is a direct transition from the present state to one or more other states, or to the same state itself in the form of a recursive loop.[2] Figure 1 is a graphical representation of the finite state generator used in the introductory Illustrative Study. The numbered circles represent the states of the system and the arrows represent the rules of transition permitted from one state to the next. From S_2, for example, the grammar permits us to move to S_5, then back again to S_2, thus: $S_2 \rightleftarrows S_5$. This loop can be traversed any number of times. The nonsense words associated with the arrows represent the finite set of symbols that comprise the given vocabulary of the language. This particular finite state system has a vocabulary of three words—NEN, PAG, and ZIR, which could be abbreviated as N, P, and Z for convenience. One of these words is generated each time we move from one state to the next. For example, in S_3 the rules say either move to S_4 by generating the word PAG, or make a recursive loop by returning to the same state itself by generating ZIR any number of times. Upon the first visit to $S_{0'}$ a sentence is completed; a sentence is any well-formed string. A well-formed string is generated by always starting in S_0 and following any route in the diagram and ending in the first visit to $S_{0'}$. The well-formed strings



RULES OF TRANSITION

(0,Z,1)	(1,P,3)	(3,Z,3)	(5,P,2)	(7,Z,8)
(0,Z,6)	(2,N,5)	(3,P,4)	(6,P,6)	(8,N,8)
(1,P,2)	(2,Z,4)	(4,N,0)	(6,N,7)	(8,P,0)

BASIC STRINGS	CANONICAL NOTATION
ZPPN	ZP (Z) PN
ZPZN	ZP (NP) ZN
ZNZP	Z(P) NZ(N) P

FIG. 1. A graphical representation of the finite state generator of sentences used in the Illustrative Study. (An admissible string is any sequence of words generated by starting in state 0 and ending on the first visit to 0′. All admissible strings can be expressed by means of the canonical notation given below the figure. Loops have been omitted in the basic strings. The rules of transition given beneath the diagram comprise an equivalent description of the generators. Each rule defines the state in which it is applicable, the symbol it generates, and the next state in which the rule leaves the system.)

---

[2] A recursive loop in a string is the unit that can occur in the string from zero to any number of times. Expressed graphically, a loop in the following string $S_0 \xrightarrow[a]{} S_1 \xrightarrow[b]{} S_2 \overset{a}{\circlearrowright} \xrightarrow[c]{} S_3$ occurs at S_2 and is indicated by the looping arrow a. A string could have any number of loops at any of the states. Each of the grammars in Experiments III and IV contains a double loop in one of its strings.

selected by the generator in Figure 1 are: ZPZN, ZPPN, ZNZP, ZPZPN, ZPNZP, ZNZNP, ZPNPZN, ZPNZNP, etc. Of these well-formed strings the first three we consider as basic strings. The important feature defining a basic string is that no state is visited more than once. When loops are inserted, states are revisited; basic strings with their loops we are naming formulas. Basic strings are significant because, as Chomsky and Miller have shown, they provide the pattern for all the longer sentences in the languages; in grammatical terms, they are the linguistic "frames" into which the elements must fit. Any number of different sentences can be generated by following the rules and using the recursive loops any number of times; if we indicate loops by parentheses, however, all the sentences in the system shown in Figure 1 can be represented by a canonical notation of the general formulae:

ZP(Z)PN
ZP(NP)ZN
Z(P)NZ(N)P

At the bottom of Figure 1 are stated the rules of transition, the basic strings, and the canonical notation for the formulas. For example, (0, Z, 1) is a rule of transition; it means that when the system is in S_0 it can generate the symbol Z and thus move to S_1. One direct path traced by a set of rules of transition, *without loops,* is a basic string: *with loops* it is a formula. Thus (0, Z, 1), (1, P, 2), (2, Z, 4), (4, N, 0′) yield the basic string ZPZN; by adding the following two rules, (2, N, 5) and (5, P, 2), a double loop has been added to the basic string which gives the formula ZP(NP)ZN.

Finite state systems provide attractive materials for experimental studies in psychology. Their formal simplicity and psychological complexity make them potentially useful for the study of higher mental processes. Moreover, the pattern and construction of finite state systems lend themselves to quantitative handling by information measures. These materials are a new, useful, and quantitative tool to apply to the study of memory and thinking. For these reasons we choose to use finite state systems as materials for exploring pattern conception. From a finite sample of sentences, subjects are expected to infer the underlying grammar. To what extent controlled variation in the formal structure of these systems may affect the behavior of the subject broadly defines the scope of the present research.

At the beginning of an experimental session a finite set of sentences from a particular language is presented to a subject in a rectangular array of cards on a table surface. The cards have been previously arranged by tables of random numbers and are all exposed simultaneously before the subject. (This procedure of presenting all the cards to the subject at once is different from the procedure followed in the Illustrative Study.) Each 3″ × 5″ card contains one string of words only; the string is typed in capital letters on a single line. The subject sits facing the cards; he reads the typed instructions, then repeats them to the experimenter in his own words.

Neutral instructions: "Here is an array of cards to be sorted into groups, into as many or as few groups as you wish to make and for whatever reason you wish to make them. When you are satisfied with your groups inform the experimenter."

If a subject has any questions regarding the time limit he is informed that there is none. Other questions are answered by merely referring him to parts of the typed instructions. When a subject begins to study the cards the experimenter starts a timer manually; the experimenter stops the timer as soon as the subject is through with his sorting. After the subject has made his groups he is asked to generate two sample strings to match each of his groups and the experimenter records the results of the subject's sortings along with any comments that the subject is willing to make about the way he performed the task.

The following principles were implicit in the procedure: (*a*) the problem is presented in as unstructured a way as possible in order to give the subject freedom to structure the materials by his own rules, (*b*) the subject is free of time pressures, and (*c*) there is a three-way check (sorting, additional instances, and free comments) on whether or not the subject was using any consistent strategy in arranging the cards into categories.

Subjects

Graduate and undergraduate students of Radcliffe College and Harvard University in the age group between 18 and 35 years, irrespective of field of specialization, were run as paid subjects. No subject served twice in any part of the experiment.

An Exploratory Study (Experiment II)

This experiment was a pilot study (*a*) to explore the kinds of strategies subjects use in grouping a given set of strings generated by finite state grammars (if, indeed, they used any strategy at all), and (*b*) to define some quantitative methods to measure the amount of "pattern" in each system of categories developed by a subject.

Strategies

Definition. A strategy may be qualitatively defined as the method or rule used by a subject for grouping a number of given strings into subsets. Operationally, a strategy is determined by the specific sentences that a subject puts into a subset. A specific strategy equates with a specific pattern a subject uses for making his groups. (This use of the term is different from the way Bruner defines it—Bruner et al., 1956, p. 54).

Quality. A strategy may be successful or less successful. It is considered successful if it corresponds completely with the basic strings of the experimenter's grammar. It is less successful if it deviates from the experimenter's rules. The degree of correspondence between the experimenter's rules and a subject's rules determines the degree of the subject's success. No strategy used by a subject is labeled as completely unsuccessful.

Procedure

The materials and procedure followed in this study have been described above; they are essentially the same in all the studies described below. In this exploratory study nine different finite state systems were set up in a 3 × 3 design; three different generators (grammar) of three, four, or five basic strings (see diagrams in Figures 2a, 2b, and 2c) were combined in all possible ways with three sizes of vocabulary, either of three, six, or nine words. Nonsense words were used in all the languages in order to minimize possible associations with the stimulus sequences. A tenth language consisted of sequences of random strings which were generated with the help of tables of random numbers. A vocabulary of three words was used for this set.

The languages were constructed with the following principles in mind: (*a*) Grammar III (five basic strings) is a hybrid composition from Grammars I and II. (*b*) The grammars are constructed to facilitate the experimenter in detecting objectively the strategy used by a subject. This principle places a constraint of congruence of certain parts—initial and/or final—in the structure of the basic strings. (*c*) The placement of symbols on transitional arrows is determined partly by the reason given previously, partly by haphazard assignment. (*d*) The number of instances (samples) in all ensembles in this study is 16. This particular design of three types of grammar and three sizes of vocabulary was used to include strategy construction under a variety of conditions to be tested more specifically later. The random strings condition was designed to serve as control for the nonrandom conditions. The nonrandom conditions are the nine grammar groups.

Sixty subjects were run individually, six subjects per condition. Subjects were distributed one per condition in the order in which they came.

Analysis of Strategies

Qualitative Analysis

There were 16 instances of sentences from each of the languages; the cards were numbered on the backs from 1 to 16. Each subject's sorting was recorded in terms of these numbers. The grouping produced by a subject was then compared with the various groupings that would

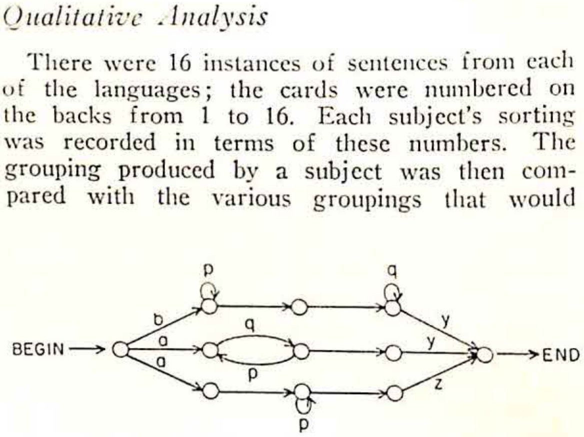


FIGURE 2a --GRAMMAR I

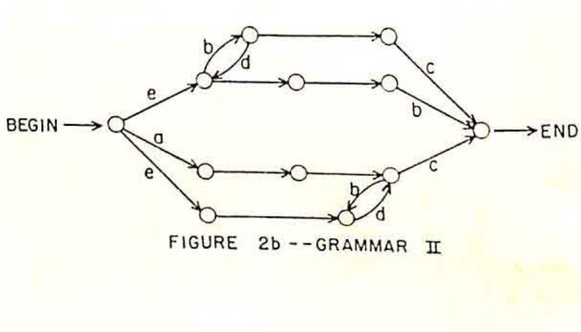


FIGURE 2b --GRAMMAR II

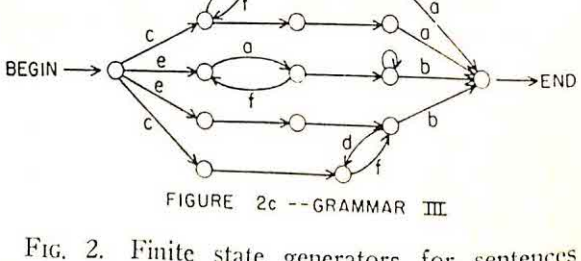


FIGURE 2c --GRAMMAR III

FIG. 2. Finite state generators for sentences used in the Exploratory Study (Experiment II). (The small letters on arrows indicate similarities and differences in the initial, final, and recursive-loop positions of symbols in the sentences generated).

result from following the different strategies. Thus, if the subject's groups were (3, 7, 15, 16) (1, 5, 10) (11, 12) (2, 6, 13, 14) (4) (8) (9), they corresponded with groups that resulted if the Length strategy were followed. Length strategy refers to grouping the instances by the number of words—the length of the sentence—contained on each card. The experimenter had worked out various groupings in terms of possible strategies so that a strategy could be objectively identified. Of course, the matching samples generated by a subject for each of his groups and the protocol were additional checks upon the identification of the subject's strategy.

Out of the billions of possible partitions of the 16 instances only about 20 strategies, not always markedly different from each other, were found to be used by subjects. The five most common strategies were labeled as follows:

Length: Subject puts together instances having the same number of words.

Alphabet: Subject alphabetizes the whole set of instances, treating it as a single category.

Initial: Subject sorts by similarity found in the beginnings of sentences. This similarity could range anywhere from one to four words.

Initial-Final: Subject sorts by the identicalness both of the beginnings and the endings of sentences.

Formula: This strategy approximates to the experimenter's rules for the finite state generator.

Miscellaneous: This category is a cover-all group for isolated cases of sorting such as sorting by a single striking word, or by the repetitive elements, or by rhythm, etc.

Quantitative Techniques

Since a variety of strategies was employed, strategies ranging from partially successful ones to the one that does succeed completely, it becomes necessary to order or scale the strategies. Earlier, we defined a successful strategy as one that completely corresponds with the basic strings of the experimenter's grammar. Three measures for scoring strategies were employed: an agreement-disagreement measure, a rating scale, and a time score. The first two were constructed specifically for evaluating strategies in this research.

Agreement-Disagreement Ratio (A-D ratio). This statistic is a kind of correlation technique for measuring the degree of agreement and disagreement between the experimenter's grammar and a subject's categories; or between any two subjects' categories. The ratio expressed (roughly) by the observed number of agreements over the total number of possible agreements in grouping, yields a measure of correspondence. When the two groupings are identical, the A-D ratio equals unity (see Appendix A for details of this technique).

Rating Scale (R scale). Each subject's strategy was assessed on a 10-point scale developed for this purpose. This scale was set up in the following way. Each subject's performance was classified into a strategy by the methods previously stated. Approximately 20 strategies seemed available to the subjects. Identical randomized lists of these were presented to five judges for rating. The judges were graduate students at Radcliffe College and Harvard University. They represented different fields of experience and cultural backgrounds. The judges were instructed jointly to rank on a 10-point scale the strategies subjects used. The instructions given to the judges were:

> You have been handed a list of labels for the strategies subjects used in an experiment in sorting into groups a pack of cards prepared by the experimenter. [Judges were shown the cards and the experiment was explained to them. Also, the labels of the strategies were described or explained.] You are being asked to rate these strategies by assigning them values between 1 and 10. Rate the poorest strategy 1 and the best one 10. You may assign the same value to several strategies if you think they rate the same and you may indicate distance in the quality of strategies by the size of the interval you leave between any two. Any questions?

After the questions were answered the judges worked independently. From the averages of the scores assigned by the judges to the strategies a standard 10-point rating scale for evaluating strategies was developed (see Appendix B for details).

Time Measure. A record of the total time required for sorting by each subject was made. Thus the number of seconds per word provided a third measure of a subject's performance.

Reliability of Measures

The A-D Ratio. Although this statistic is a fair measure of success in terms of the correspondence of a subject's responses to the experimenter's rules for generating the strings, it is subject to experimental artifact. For example, the Initial$_1$ strategy (instances grouped together on the basis of the first word only) yields a spuriously high correlation score. The number of subjects using the Initial$_1$ strategy is comparatively small, however, and the other two measures offer a second check on the results.

A second difficulty with this statistic is that interlanguage comparisons cannot be made without using some kind of correction factor for the number of instances in the set. The measure of agreement rises with the increase in the number of instances; the measure is not really comparable from one language system to another.

The R Scale. The following table shows the reliability of the ratings of strategies made by five judges. Rank-difference correlations were used.

It should be noted that the equality of intervals between ranks in this scale cannot be established.

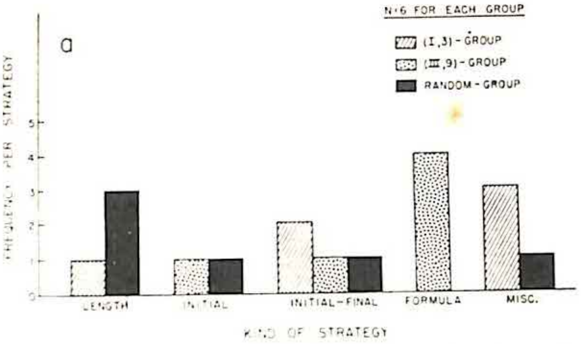
TABLE 1

RANK-ORDER CORRELATIONS BETWEEN JUDGES FOR RATING STRATEGIES

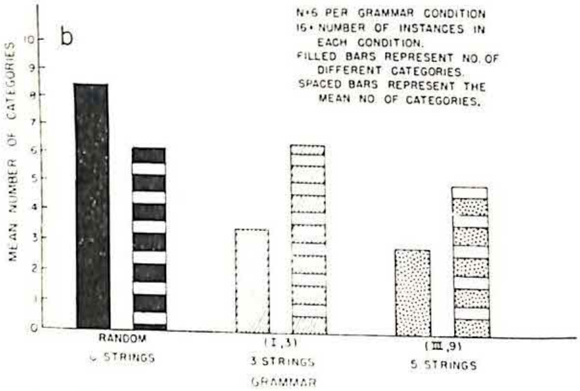
Judges	1	2	3	4
2	.89			
3	.94	.77		
4	.88	.89	.82	
5	.91	.87	.89	.88

Note.—From the ratings assigned by these judges to strategies was developed the 10-point rating scale.

Time Score. Although time measures have been commonly used in studies of visual pattern perception (Adams, Fitts, Rappaport, & Weinstein, 1954; Anderson & Leonard, 1958; Fitts et al., 1956), in the present research they are only a rough measure. Several factors are simultaneously operative which decrease the reliability of the time scores. As the number of instances in the set increases, the time for sorting increases disproportionately, paralleling the rule that time increases according to the amount of information that must be processed. Both length and complexity of task tend to increase time, but if a subject works with the self-instruction of speed, he finishes the task much more quickly. The difficulty here is circular. The less time a subject spends with the problem the less information he extracts from the input and the less time he spends dealing with his problem.



3a. Frequency comparison of strategies by subjects in the Random and two grammar groups. (The Random group worked with 16 strings that had been haphazardly generated. The (I,3) group worked with 16 well-formed strings that had been generated with three formulas and a vocabulary of three words. The (III,9) group worked with 16 well-formed strings that had been generated with five formulas and a vocabulary of nine different words.)



3b. Mean number of categories for subjects in the Random and two grammar groups. (Solid columns—first in each pair—compare the same three languages for the number of different categories made. Different categories refer to instance categories, e.g., one-instance category, two-instance category. Broken bars compare the same three groups for the average number of categories for each group.)

Fig. 3. Frequency comparison of strategies and mean number of categories.

Results

The distribution of subjects in the Random group is compared in Figure 3a with two of the nine groups (I,3 and III,9) with respect to the kinds of strategies used in each. The bars for the Random group and the (III,9) group (five basic strings and a vocabulary of nine words) are longer at the opposite ends. Condition (I,3) (three basic strings and a vocabulary of three words) has more affinity with the Random group than is apparent from the figure; three of its subjects who were classified under the Miscellaneous condition in the type of strategy used actually used some kind of differentiated counting technique that was slightly more complex than the plain counting involved in the Length strategy. Figure 3b compares the same three groups by the average number of different categories made by each in the first columns (filled bars). (By different categories we mean instance categories; for example, one-instance category, two-instance category, and so on.) Notice how the Random condition, where every instance belongs to a separate category, averages to the largest number of different categories. The second part (spaced bars) compares the groups for the mean number of categories made by each. Here the difference between the Random group and (I,3) group is negligible.

Table 2 is a frequency tabulation of all subjects showing their distribution by the kinds of strategy they employed and the experimental conditions under which they had worked. Note that there are just about five or six types of strategy that cover all subjects under the various conditions tested. Also note that the better or the more successful strategies are used under some conditions more than in some others. For example, none of the 18 subjects used the Length or the Alphabet strategies when the number of alternative words used was nine. Ten of these subjects used some form of the Formula strategy. When the sentences were formed with a vocabulary of only three words, on the other hand, 5 of the 18 subjects used the simpler strategies and only 4 discovered the Formula strategy; all of those 4 subjects came from the same condition group.

For each of the 10 experimental conditions, Table 3 presents means which have been calculated by assigning values to strategies by the statistical techniques discussed above. Because the experimenter worked with no rules in the Random set no A-D ratios can be calculated for that strategy. Ratios for Condition III are found to be higher for each constant condition of vocabulary size, except in the grammar Group (II,3), which may be treated as an exception in all the tables given subsequently because of an artifact in the construction of the language. The group means for the R scores are tabulated in the second row. Here the Random group and (I,3) group are more alike, as we have com-

TABLE 2

FREQUENCY DISTRIBUTION OF SUBJECTS IN THE RESPECTIVE LANGUAGE GROUPS BY THE TYPE OF STRATEGY THEY USED

Grammar	Kind of Strategy						N
	Length	Alphabet	Initial	Init.-Final	Formula	Misc.	
Random$_3$	3		1	1		1	6
(I,3)	1			2		3	6
(II,3)	1	1			4[a]		
(III,3)	1	1	1	3			
Total	(3)	(2)	(1)	(5)	(4)[a]	(3)	18
(I,6)	2		1	1	1	1	
(II,6)	2		2		2		
(III,6)			2	2	1	1	
Total	(4)	—	(5)	(3)	(4)	(2)	18
(I,9)				2	4		
(II,9)			3	1	2		
(III,9)			1	1	4		
Total	—	—	(4)	(4)	(10)		18
Sum Total	10	2	11	13	18	6	60

Note.—Figures in parentheses are sums of frequencies immediately preceding those totals.

[a] These four subjects all belong to the (II,3) group.

TABLE 3

Group Means for Ten Experimental Conditions

Mean Score	Language Group									
	0,3	I,3	II,3	III,3	I,6	II,6	III,6	I,9	II,9	III,9
A-D ratios	—	.46	.773	.616	.437	.601	.73	.725	.805	.85
R score	1.83	2.83	7.0	3.71	3.66	4.66	3.33	6.83	5.5	7.2
Time score	6.04	6.52	4.04	6.36	3.61	4.75	4.64	4.09	.44	6.05

Note.—The means have been calculated by three independent statistical measures: A-D ratios, Rating scores, and Time scores. Time scores represent average time taken per syllable.

mented before, and both are inferior to Groups (I,9), (II,9), and (III,9). The time comparisons in the third row are somewhat fuzzy. Sometimes a subject, who may have worked with the instances for a very long time, will settle on a very poor strategy at the end; sometimes, the opposite is true.

A comparison of the group means by vocabulary and by grammar by applying all three measures is made in Table 4. Although the figures given in this table are not always directly comparable, their relative increase or decrease with change in the experimental condition is constant.

Discussion

This first study was intended to explore the kinds of responses subjects make to a categorizing task. The task as we have defined it is one of sorting strings of words simultaneously presented to a subject on small cards. The subject is free to work on these instances for as long as he wishes, without receiving further help from the experimenter, however, or from the instructions that are given to him. He is neither told how many categories he should make, nor is he given any indication of success or failure. The criterion of success, if any, is set by the subject himself and he does not have to form groups if he thinks no groups can be made.

Though the experimenter is somewhat in the dark as to what the nature of the controlling variables is, he does, however, anticipate a few problems. Therefore, he sets up his first study under a variety of experimental conditions to get leads about the possible functional relations between some of the experimental variables and subjects' responses. The basic variables that the experimenter works with are: the number of formulas which provide the

TABLE 4

A Comparison of Relative Difficulty of Grammar and of Vocabulary as Indicated by Three Independent Measures

Grammar:	Random	I	II	III
A-D ratios	—	.53	.72	.73
R score	1.83	4.44	5.7	5.1
Time/Word	6.04	4.85	5.91	5.59
Vocabulary:	Random	3	6	9
A-D ratios	—	.62	.59	.79
R score	1.83	4.6	3.9	6.6
Time/Word	6.04	5.48	4.51	6.24

basic framework for the strings and the number of different alternative words with which he develops his formulas.

The results illustrate several important features about subjects' responses in an information processing situation. First, subjects process information by rules, or as we have suggested earlier, by pattern conception. Next, we note that the rules formulated by more than 90% of subjects can be classified under five headings: Length, Alphabet, Initial, Initial-Final, and Formula strategies. Third, there is a hierarchy within these five or six types of strategies which has been identified in subjects' reponses. These strategies vary from simple to complex. Finally, the results indicate that some experimental conditions are better than others for inducing the formulation of the more successful strategies. The reader will recall that the more successful strategies have been defined as those that have a higher agreement with the experimenter's formulas. The Random strings, which were haphazardly constructed, embody a condition which is least favorable to the emergence or acceptance of pattern strategies in subjects' behavior. Evidence from our data suggested that vocabulary and grammar may be two of the more effective parameters in pattern conception.

Pattern Conception. What we had observed as a dramatic instance of orderly responses demonstrated in the introductory study, was confirmed by the responses of more than 50 subjects in this exploratory experiment. It is not a fortuitous matter that no subject, not even one from the Random group, handed the cards back to the experimenter in some haphazard arrangement or declared the task insoluble. On the contrary, in each case the subject used some kind of a system by which he had looked for relations among the instances to create his categories. Even with the Random group some subjects worked very hard and long trying to find some sort of scheme to organize their information. By and large, the subject could verbalize the type of relations that he had discovered in terms of a rule or rules though he was not always able to do this with precision.

Hierarchy in Rules. The thing that amazed us most was that nearly all subjects could be classified so easily within the same general types of five or six strategies—strategies which were, by and large, the same ones we noted earlier in the responses of our single subject described above in the Illustrative Study. In other words, a hierarchy in rules that was indicated in the individual's responses seems to characterize the group responses as well. This is a two-sided observation: it points first, to the fact that responses under all conditions of card sorting described above are classifiable within five or six easily detectable strategies; and second, that these strategies themselves can be ordered in complexity.

We can classify these strategies into two types: the ready-made type and the invented type. Ready-made rules are independent of the nature of the information to be processed. The rule of counting, for instance, to make groups of equal length, already exists in a subject's repertory and can be easily transferred to almost any kind of material. The ready-made rule applies whether the materials be heads of cattle, pictures in an art gallery, or the coded messages about the enemy's moves. This rule of counting is an instance of a simple quantitative strategy that is highly useful in dealing with certain kinds of information, but not so appropriate for certain other kinds. The counting rule may take on slightly different shades, as counting the number of different words, or counting the number of repetitions of any word, etc., but essentially it is the same rule that is applied.

Alphabetizing is here classified as a ready-made strategy because this rule, too, already exists in the everyday repertoire of a college student. Since it is always possible to alphabetize a given series of words of any language, it requires little ingenuity on the part of a subject to apply this rule.

It is important to note that a subject loses information by applying the simple strategies of counting and alphabetizing to these not-so-simple sequences. We have, therefore, evaluated such strategies as being less successful. A strategy is rated as less successful in proportion to the amount it deviates from the finite state grammar which governs the given strings. In the simpler strategies a subject seems to work more with the perceptual characteristics of the information than with the structural ones.

Subjects in the Random group, which is the control group for all the nine experimental groups, seem to resort to the simple strategies, particularly the one of counting, more than the subjects in the grammar groups seem to do, except the subjects of the (I,3) group. The responses of this group have close kinship with the responses of the Random group. When a subject in any group either finds the task too difficult for him or when he does not wish to exert himself, functionally, he is treating the grammar strings as if they were random strings. Hence the strategies that subjects would be found to be using in these instances would be close kins to counting or to alphabetizing strategies. It seems highly probable that subjects in the (I,3) group were treating the grammar strings as random strings.

The observed fact that a subject in the Random group follows a counting strategy does not necessarily imply that the subject has worked with one strategy only. Subjects' protocols bear evidence to the contrary. Several subjects described complex strategies that they had tried to apply to the strings, but which they had discarded because they always seemed to find too large a number of single cases. Notice that a subject is uneasy about having too many one-case categories, and this is virtually what he would find in the random set where each instance is unique. After having searched for abstract relations, a subject is forced by the stimulus properties of the set to settle for something less. This sample of behavior is apropos to the basic argument in this thesis.

Strategies labeled Invented are different from the ones just described. We classify the Initial_n-Final_n strategy and the Formula strategy under this label because it is the nature of the materials which suggests these strategies to the observer. A high degree of discovery, however, is also involved in these strategies, but since a subject is unaware of the embeddedness of any formulas within the materials, we prefer to call them the Invented rules rather than the Discovered rules. This label, however, applies most appropriately to the Formula strategy. In this strategy the subject has to go beyond the obvious information. The subject has to tease out an invariant sequence of words in each of the instances, as also he must find the recursive units. The search for invariance and variance yields formulas which are the basic framework for the grammar strings. This type of response calls for a certain degree of abstraction in order to discover the true pattern implicit in the strings.

Invented strategies correlate more closely with the information given. Since the instructions a subject receives are neutral, his only source of information is the set of strings he has before him. The amount of information he is able to extract from them is a critical factor in the way he will process the information. His efficiency in extracting the maximum amount of useful information depends largely upon two, sometimes interdependent, conditions of motivation and stimulus complexity.

Motivation. Each subject seems to redefine the problem to himself and it is very important how he does define the task. Subjects who look for quick and easy solution—whose motivation appears to be to escape the situation as quickly as they can —do not get themselves involved in the real problem of finding relations among sets of instances. They tend to manipulate the more obvious characteristics of the stimuli and fail to extract useful information contained in the sequence order of the words. Though they might also see other little bits of isolated information, they tend to ignore these and proceed to apply the easy rules of counting and alphabetizing, thus minimizing the cognitive strain.

Stimulus Complexity. Subjects who do not get seriously involved in the task may or may not end with the Invented strategies. One subject in the Random group worked with her 16 cards for 41 minutes. She tried to make groups by initial similarities, by ratio relations, by sequence relations, and by a method of substitution. But when none of these strategies worked by her criterion of success, she resorted, according to her own version, to "the simplest strategy of counting." The task of seeing relations among the random strings is perhaps psychologically more complex than the task of seeing relations where relations do in fact exist.

The results from all nine grammar group conditions of finite state systems where relations do in fact exist, do not distribute under the Invented strategies only. Some do and some do not. In Table 2 we notice that the stimulus condition, where nine alternative words were used to develop the strings of sequences, as compared to three in the Random group, and three and six in the other two conditions, more than 50% of the subjects evolved the Formula strategy as compared to the 0% of the Random group and 22% in each of the other two. Of these, the 20% of subjects in the vocabulary three group, all come from a single condition, (II,3), meaning that none of the subjects from Condition (I,3) or (III,3) or (0,3) arrived at the Formula strategy.

Are these differences, indicated in the percentages of subjects who evolve the Invented strategies and those who do not, functionally related to the complexity of the experimental variables under which subjects work? If so, what constitutes experimental complexity? That is the question we turn to in the next section.

Experimental Variables

In the exploratory study of the previous section evidence for a search for pattern was found. Sometimes this search was creative, sometimes transitive, but what was not clear from the data was why this difference existed in the quality of performance. Informal evidence suggested that instruction was one variable that needed to be controlled more precisely and the data indicated that some stimulus features of the input were contributing to the outcome. Accordingly, some experiments were carried out to assess the effect of the following four variables:

1. Instructions. Four different instructions were tested to select one that would maximally involve the subject in the task, but with minimal guidance from the nature of the instructions.

2. Number of instances. The number of instances per formula was varied from two instances to three to four, to test the effect of the amount of evidence on the quality of strategy.

3. Number of different alternatives (Vocabulary). Three or nine different words comprised the vocabulary for developing the different language systems.

4. Structural similarity (Grammar). The effect of the formal structure of the languages, three rules as against five rules with variation in similarity between the rules, was also investigated.

The Instructions
(Experiment III)

Materials

Two new artificial languages, $(I,9,16)_b$ and $(III,3,23)_c$ were constructed to study the effects of four types of instruction on each. Language $(I,9,16)_b$ is constructed with three formulas in its grammar; its vocabulary consists of nine words. Sixteen instances were selected to make up this set. Language $(III,3,23)_c$ contains five formulas; three words have been used for its vocabulary. This set contains 23 instances of well-formed strings. Two additional language conditions, (I,3,16) and (I,3,23), were also tested, both under "eliminate speed" instructions. Altogether, then, there were 10 language conditions that were compared in this study.

Procedure

The same general procedure described earlier was used in all experiments reported in this section. The specific variable to be considered here is the kind of instruction to be used. Ten groups of subjects, four subjects per condition, were run under four different instructions. No subject served twice in any part of the experiment. Subjects were distributed, one per condition, in the order in which they came. The following four sets of instructions were used:

1. Neutral. "Here is an array of cards to be sorted into groups, into as many or as few groups as you wish to make, and for whatever reason you wish to make them. When you are satisfied with your groups inform the experimenter."

2. Eliminate Speed. "Here is an array of cards to be sorted into groups, into as many or as few groups as you wish to make, and for whatever reason you wish to make them. Your performance will be judged *not* by speed, that is, not how quickly you are through, but by *what* you do. When you are satisfied with your groups inform the experimenter."

3. Relations. "Here is an array of cards to be sorted into groups, into as many or as few groups as you wish to make on the basis of some relations that you can find in the members of the groups. Your performance will be judged *not* by speed, that is, not by how quickly you are through, but by how well you discover relationships. When you are satisfied with your groups inform the experimenter."

4. Logical. "Here is an array of cards to be sorted into groups, into as many or as few groups as you wish to make on some logical basis. Your performance will be judged *not* by speed, that is, not by how quickly you are through, but by how logical you are. When you are satisfied with your groups inform the experimenter."

Results

Table 5 presents the means, ranges, and critical values for the A-D ratios, R scores, and Time scores for 40 different subjects run under 10 different experimental conditions. The critical differences between groups were calculated by the Link-Wallace method of short-cut allowance for one-way analysis of variance (Mosteller & Bush, 1954, p. 305). This statistic uses the range as a measure of variation instead of the sum of squares.

Conditions F, G, H, and I in Table 5 apparently belong together. The difference between any two totals in this group is below 1.51 in the A-D ratio comparisons, below 18.80 in the R score comparisons, and below 41.59 in the time comparisons. It is clear, then, that variation in instructions fails to produce any recognizable difference in the performance of subjects in language (III,3,23)$_c$. Presumably, the task involved in this language is too hard for the subjects. Even when the subjects (Column H) took longer than subjects in any other group, they failed to give efficient results. Since the complexity of this language masks the effect of instruction (if any), we shall concentrate on the simpler Grammar I.

Conditions C, D, E, and J likewise obviously belong together. None of the differences between any two of these con-

TABLE 5

MEANS, RANGES, AND CRITICAL VALUES

Language	Instructions	A-D ratios			R score			Time score		
		M	*R*	*CV*	*M*	*R*	*CV*	*M*	*R*	*CV*
A (I, 3, 16)	Elim. Speed	.333	.40		1.75	3		2.38	1.39	
B (I, 9, 16)$_b$	Neutral	.402	.433	1.95	4.25	7	0.47	3.55	3.56	41.58
C	Elim. Speed	.69	.25		7.5	6		3.37	2.25	
D	Relations	.95	.20		9.75	1		5.49	4.35	
E	Logical	.76	.14		7.5	6		5.35	2.71	
F (III, 3, 23)$_c$	Neutral	.58	.21	1.51	4.0	2	18.80	7.81	16.94	41.59
G	Elim. Speed	.50	.43		5.25	5		9.04	5.99	
H	Relations	.58	.27		5.0	4		18.58	22.59	
I	Logical	.49	.43		5.75	5		8.55	17.70	
J (I, 3, 23)	Elim. Speed	.79	.46		8.75	1		9.76	10.09	

ditions exceeds the calculated critical values in any of the three measures. All of these four conditions belong under the simpler Grammar I, which also includes Conditions A and B. Subjects who served under Condition A performed most poorly, even though, as in C, D, E, and J, they were instructed not to speed. Subjects for $(I,9,16)_b$ who served under Neutral instructions (see Column B) do not resemble or differ consistently with the other groups.

Discussion

In the preceding section we raised a question about why some subjects follow easy strategies and some others follow more complex ones. An analysis of the subjects' protocols gave an indication that one necessary, though not sufficient, condition for evolving a good strategy is that the subject should extract all the *useful* information from the instances he receives. This business of extracting the most useful information is in itself a very critical step in pattern conception. The success with which a subject decodes relevant information from the stimuli depends partially upon the degree to which he involves himself in the task. This in turn is influenced by a variety of factors, such as the past history of the subject, his present motivation, his age, etc., as well as the instruction he receives from the experimenter regarding the nature of the problem, other things being equal. An individual's history or his motivation is difficult to control by laboratory techniques, but the experimenter can explore the instruction variable.

Actually, a subject receives two kinds of information from the experimenter. One kind comprises the main body of information that is contained in the strings of nonsense words with which the subject is supposed to work; the other is some information *about* this main body of information to be categorized. The main body of information is contained in what Bartlett (1958) has called a "closed system," which the subject must discover. The efficiency with which a subject processes the main body of information will in part depend on how well he has decoded the information from the instances presented to him. It is the experimenter's wish to get a subject started on this search but without telling him how to proceed. Thus in setting up the instructions the experimenter keeps two things in mind: to get a subject maximally involved in solving the problem and to do this with a minimum of guidance from outside.

Neutral instructions are suspect when they are evaluated by this two-edged criterion. In the absence of specificity, both of instruction and of materials (nonsense strings), a subject may often resort to self-instruction, either to work for speed, or to find some easy-way-out strategy; he thereby loses much information. To overcome this difficulty it seems necessary to furnish a subject with some other kind of instruction.

French (1954a) has shown that performance improves with more specific instructions in a task that involves sequential dependencies. The question here, however, is not one of improving a subject's performance by giving him more guidance or by making the instructions more specific—we take that for granted—but rather the question is one of finding the kind of instructions that will induce a subject to make the maximum effort to solve the problem, and to bring this about with as unstructured a kind of working instruction as possible.

We noted in the results that when a language is too difficult for the subjects, as $(III,3,23)_c$ seemed to be, the effect of instructions is confounded. But with the simpler Grammar I, subjects do better when they are instructed to slow down, with the one notable exception of (I,3,16) —see Condition A. We pointed out earlier that the results of Condition A are the poorest of all the conditions tested. A differs from B, C, D, and E in the size of the vocabulary; it differs from F, G, H, and I in its grammatical structure; and its difference from Condition J seems to be clearly related to the number of instances. Presumably, A suffers in part from one or more of these conditions.

The difference between B and Conditions C, D, E, and J was attributed to different instructions. All four conditions, C, D, E, and J, required subjects not to rush (note that Relations and Logical also eliminate speed), and they all emphasized that it is *what* a subject does that is more important than speed. Condition B is neutral to these emphases. It does not seem unreasonable, therefore, to suppose that the emphasis on "what you do" increases the search for useful information, and this search for the useful information, we have granted, is one critical step in pattern conception.

Johnson (1955) discusses the effect of instruction on performance under the concept of "set" (see pp. 64-67). After defining set as "temporary integration of behavior," he elaborates on its meaning as integration in terms of the obvious and simple things, or integration in terms of "subtle aspects of complex stimulus patterns." Explained in this way, set seems to be no different from recoding of information. Verbal instructions can drastically change the quality of recoding. As for instance, in the Carmichael, Hogan, and Walter (1932) experiment the same figures are recoded as dumbbells by one group and as spectacles by another group of subjects, depending upon the instructions the group received. In studies such as this one, the experimenter has already laid down the recoding scheme for his subjects through the instructions he provides, and then he is surprised (and pleased perhaps!) to find that subjects have followed his scheme.

It is this kind of coaching in the recoding scheme of a subject that we wished to avoid in our instruction. For, after all, a major interest in the present research is to study those factors that affect the recoding scheme of a subject. Very little new information will be gained by recovering the same coins that we originally fed into our slot machine.

So, what do we finally conclude? On the whole one could say that the instructions, to the extent to which they were varied, did not produce a strong effect on the quality of the performance. Other variables (complexity of grammar, number of different alternatives, number of instances) seem to be more important because they can make the problem so hard that they mask out the effects of instructions. When conditions are optimal for instructions to have an effect, the main effect that changed instructions produced was to slow subjects down to make them stop, look, and think more carefully, and this is the only effect the experimenter wished to induce through instructions. Any more help than just maximally motivating subjects to think would be a violation of what has been described as good instruction. Since the instruction to eliminate speed, together with the emphasis on "what you do," is common to all three sets of instructions tested, it seems reasonable to suppose that this amount of change in the neutral instructions was enough to involve the subject in the task. Therefore, we adopted Eliminate Speed instructions as standard for all subsequent studies.

The Number of Instances
(Experiment IV)

Materials and Procedure

One important characteristic of an ensemble of instances is its size. The size of the set can be controlled by controlling the number of formulas of a grammar and/or the number of instances given for each formula. Moreover, one might like to control the total number of words in the ensemble. The structure of the languages used in this study is clarified in Table 6.

The Random condition contained no formal grammar; it served as the control group for all the other conditions. Languages B, C, and D are identical except for an orderly increase from two instances per loop in (I,3,13) to three instances in (I,3,18) to four in (I,3,23). The number of instances per loop and the size of the set are interrelated; as instances increase in number, the size of the set increases also.

Languages (I,3,13) and (I,9,13) are also identical except for the difference in the size of the vocabulary: B is developed with three words; E with nine. Thirteen instances for Grammar I have been considered to comprise a minimum set by definition. Condition F, $(I,9,16)_b$, represents a random increase in the number of instances from 13 to 16. The choice of the size, 16, was made to match the size of the Random set. Conditions E and F will be compared for a small

TABLE 6

STRUCTURAL ANALYSIS OF LANGUAGES

Language	Grammar or number of formulas	Vocabulary or number of different words	Instances per loop	Size of set	Total number of words
A Random (0,3,13)	0	3	1	16	118
B (I,3,13)	3	3	2	13	90
C (I,3,18)	3[a]	3	3	18	180
D (I,3,23)	3[a]	3	4	23	180
E (I,9,13)	3[a]	9	2	13	90
F (I,9,16)$_b$	3[a]	9	Irregular	16	118
G (III,3,23)$_c$	5	3	2	23	180
H (III,9,23)$_c$	5[b]	9	2	23	180

[a] Same as in B.
[b] Same as in G.

increase in the number of instances from 13 to 16, other things being held constant.

A similar comparison in size is envisaged for Languages (III,3,23)$_c$ and (III,9,23)$_c$. Grammar III consists of five formulas as against three of Grammar I. A minimum set for Grammar III consists of 23 instances, which also allows 2 instances per loop. All experimental conditions for these two sets were held constant, except for the size of their vocabulary.

One could also compare (I,3,13) and (III, 3,23)$_c$ with each other as minimum sets for their respective grammars. Since the number of instances per loop has been held constant, any differences found between the two languages would be attributable either to the number of formulas in the grammar or to the number of instances in the sets, or to both.

Ten subjects per condition, using the instructions aimed to eliminate speed, were run for each of the eight language conditions.

Results

Table 7 presents the means and critical values for 10 language conditions by the usual three measures. Only 8 of these 10 groups, A through H, concern us in this section. We note first of all that in Grammar I, Conditions B, C, and D, as the instances per loop increase from two to three to four, strategies improve correspondingly. In working with 13 instances a subject spends about the same amount of average time per word as a subject working with the larger sets of 18 or 23 instances, but he does not achieve better results. In fact, the strategies used under this condition are poorer than most and about as poor as the ones used by the Random group, which used the poorest strategies.

TABLE 7

MEANS, RANGES, AND CRITICAL VALUES

Language Group	A-D ratio	R score		Time score	
	M	*M*	*R*	*M*	*R*
A Random (0, 3, 16)	—	2.8	6	11.38	49.24
B (I, 3, 13)	.253	3.1	7	8.14	29.93
C (I, 3, 18)	.389	4.6	7	6.71	23.02
D (I, 3, 23)	.598	6.5	8	7.73	14.27
E (I, 9, 13)	.532	6.7	6	4.996	8.29
F (I, 9, 16)$_b$	.748	8.2	7	3.916	4.23
G (III, 3, 23)$_c$	.419	4.7	8	5.396	7.31
H (III, 9, 23)$_c$	.706	6.0	7	5.035	6.53
I (III, 3, 23)$_d$	.541	3.5	8	4.998	11.05
J (III, 9, 23)$_d$	.81	7.1	7	5.478	14.39
Critical Value		33.37		79.08	

Next we note that without any increase whatsoever in the number of instances, if we merely changed the size of the vocabulary from three words to nine, performance is drastically modified (cf. Column B with E and G with H). Now look at the time measures. Subjects in Column F take far less time than subjects in Column B and yet they do far better than subjects in Column B. The same is again true, though not quite to the same degree, of subjects in $(III,3,23)_c$ and $(III,9,23)_c$. Apparently, one must look for some other factor(s), instead of the number of instances, to account for these differences between the groups.

A further point of interest may be noted from a comparison of Languages (I,9,13) and $(I,9,16)_b$. Earlier we explained that the two languages are identical in every respect except for a fortuitous increase in the number of instances from 13 in Condition E to 16 instances in F; the results, however, that follow this change are not so fortuitous. Even a small random increase in the number of instances in a language with the larger vocabulary apparently makes some difference in the subjective redundancy of the set for the subjects. But when we compare Columns A and F, both sets with 16 instances in each, the Random set scores are much inferior.

Now compare the minimum sets for both grammars. We consider 13 instances for Grammar I and 23 instances for Grammar III, as minimum sets for each, by definition. (I,3,13) is not too different from $(III,3,23)_c$ nor is (I,9,13) very different from $(III,9,23)_c$. Presumably, the effect of the size of the set is either negligible or it is masked by the effects of other variables.

To summarize our observations, three points seem rather clear: the larger vocabulary decreases difficulty, more instances per formula also decrease difficulty, the effect of the total number of instances in a set is easily modulated by other variables.

It is not clear, however, as to what effect an increase in the number of formulas produces upon subjective difficulty.

Discussion

Two of the four points summarized above pertain to the number of instances in the set; they are not unrelated points. The total number of instances in a set is automatically increased with each new addition of an instance for any of its formulas. This means that two things happen for a subject. First, every single addition of an instance to the set increases the difficulty of the set by increasing the number of relations the subject must now study (if n cards have m relations, then $n+1$ cards will have $m+n$ relations). Second, every addition of an instance for any of the formulas contained in the set simultaneously reduces difficulty by increasing the amount of evidence available. This latter generalization, however, is not true for a random set of strings; an increase in the number of instances in the Random set does not reduce difficulty, but rather increases confusion.

In the context of the present discussion we define redundant strings as those that have been derived from the same grammatical formulas. They repeat information by replicating formulas or instances or both. Random strings, on the other hand, contain no formulas or any orderly repetition of information. They could be viewed as containing as many formulas as there are instances in the set, as was suggested previously; as such, they will contain only one instance per formula. A redundant set which has been generated with one, two, or more formulas will contain instances which repeat information about the formulas. Thus the number of instances per formula and the number of formulas that constitute the grammar are somewhat interdependent features in a language.

Random strings are confusing to a subject because they are difficult to package. They contain more information than any redundant language of comparable size contains (see Columns A and F) and this mass of disjointed information is difficult for a subject to handle. The larger the size of the Random set the more difficult it is for a subject to process the information it contains. Thus, the critical factor for increasing the efficiency of performance is not one of increasing the amount of haphazard information, but it is one of receiving a greater amount of more *useful* information. Reiss (1958) came to the same conclusion with his subjects when he varied the amount of initial information presented.

In the absence of repetition (redundancy) of information, relations between instances are few and far between. The subject, who is usually disposed to sorting his environment by searching for redundancy, is nonplused in a random set-up. Now introduce some redundancy into the situation by presenting to the subject two instances per formula, instead of one instance or none, and performance improves. Bartlett (1958) made this same observation on his subjects in tasks that required gap-filling by methods of interpolation and extrapolation. In the case of a simple formula, a string with one loop only, if one now gives the subject just one more instance, making it three instances per loop, the improvement is phenomenal. The subject's response almost begins to approximate to the grammatical structure of the language. One is led to think that depending upon the amount of complexity of the formulas, one could expect an upper ceiling in improvement in performance beyond which an increase in the number of instances ceases to be effective.

When the number of instances per loop is small, say zero or one, it is more possible to confuse the instances of one formula with the instances of another on the basis of superficial relations—what Wertheimer (1945) calls "and-summative aggregations" (p. 190). Actually in the ideal situation, where instances have been grouped according to the formulas they exemplify, instances which are positive for one formula are negative for all other formulas in the set. This, however, is true only when a subject has clearly seen the structure of the formulas in abstraction. Pattern conception requires the observer to make a two-fold response to the stimulus ensemble; a response of the ingroup type and of the outgroup type. The positive instances for a formula the subject must "aggregate" as members of the ingroup for that subset; all other instances become negative instances for this formula and which he must "segregate" as the outgroup members belonging somewhere else. In effect, the subject is making an inductive-deductive response. Hypothesis making and hypothesis testing are going on simultaneously as a subject decides which instances are positive and which are negative for the particular category he is working with. His certainty increases as the number of instances that fit his hypothesis increases. More instances per formula increase evidence and consequently decrease the probability of error. It is a well-known fact in communication engineering that one repeats a message as a safety device against "noise" in the channel. The noise which the negative instances create for a subject in our channel can be reduced by increasing the number of positive instances, other things being equal.

Thus we are led to conclude that within certain limits, the number of instances per formula reduces difficulty, subject to certain interaction effects from other variables. Although an increase in the size of the ensemble, within a reasonable range, increases the difficulty of the set, the quantitative aspect of the set is relatively unimportant. If, however, the size of the set is too large the problem is either rendered insoluble, or it becomes ridiculous, depending upon the number of basic strings in the set.

The interaction effects we referred to above have pointed in the direction of the size of the vocabulary and of the complexity of grammar as two possible variables that affect pattern conception. In fact, one of the points we noted from our results was that a large number of different words reduced the difficulty of the task for the subject. We could not see our way clearly on the question of grammatical complexity. Let us explore the effects of these two variables further.

The Number of Different Alternatives and Structural Similarity

(Experiment V)

Two variables that have been clouding the analysis of our data are the parameters of the number of different words and of structural similarity of the strings. We have equated the number of different alternatives with the vocabulary of a language, and its grammar, roughly, with structural similarity. Two sets of alternatives, consisting of three words and of nine words, constitute the size of the vocabulary of the languages to be tested. The generators for these languages consist of three basic strings, five basic strings with replication and five basic strings without replication. The Random set of strings has no grammar, of course.

Materials

Two additional languages, (III,d), with vocabularies of three and nine words, respectively, were constructed to control similarity in the formal structure of the grammars. By similarity in structure (in this context) is meant: (*a*) One grammar contains some information about the other grammars. For instance, Grammar (III,d) contains five formulas, three of which are identical with the three formulas of Grammar I, and with three of the five formulas of (III,c). (*b*) Structural similarity also means that one formula contains some information about the other formulas. For example, Formulas 4 and 5 in Grammar (III,c) repeat Formulas 2 and 3; in (III,d) they do not. This explanation of the structural

TABLE 8

t VALUES FOR A-D RATIOS AND R SCORES

		A-D ratio									
		Random	(I,3,13)	(I,3,18)	(I,3,23)	(I,9,13)	(I,9,16)$_b$	(III,3,23)$_c$	(III,9,23)$_c$	(III,3,23)$_d$	(III,9,23)$_d$
R score		A	B	C	D	E	F	G	H	I	J
Random	A										
(I,3,13)	B	.03		1.97	3.35	3.00	7.35	.84	5.27	3.69	7.15
(I,3,18)	C	1.75	1.15		2.09	1.60	5.78	.16	3.88	2.41	5.77
(I,3,23)	D	2.23	2.76	1.40		.56	1.53	.90	.96	.57	2.00
(I,9,13)	E	3.93	3.43	1.8	.15		2.45	.58	1.69	.10	2.88
(I,9,16)$_b$	F	6.28	5.93	3.49	1.45	1.51		1.79	.52	3.35	.86
(III,3,23)$_c$	G	2.02	1.55	.04	1.42	1.90	3.60		1.51	.66	2.06
(III,9,23)$_c$	H	2.66	2.50	1.10	.36	.57	1.99	1.09		3.25	1.16
(III,3,23)$_d$	I	.07	.03	1.90	2.35	2.89	4.74	1.10	2.08		3.68
(III,9,23)$_d$	J	3.61	3.30	.15	.42	.31	.94	1.90	.81	2.85	

Note.—*t* values for A-D ratios are entered in upper triangle. R score *t* values are entered in left triangle. Since A-D ratios were not directly comparable for different sized sets, A-D *t* values apply for sets of equal size.

p =	.1	.05	.02	.01	.001	df = 18
t =	1.734	2.101	2.552	2.878	3.922	

similarity in the grammars can be schematized thus:

Grammar I	Grammar (III,c)	Grammar (III,d)
1. - (-) - - -	1. - (-) - - -	1. - (-) - - -
2. - - (- -) - -	2. - - (- -) - -	2. - - (- -) - -
3. - (-) - - (-) -	3. - (-) - - (-) -	3. - (-) - - (-) -
	4. - - (- -) - -	4. - - - ((-)(-)) -
	5. - (-) - - (-) -	5. - (-) - (-) - (-) -

Results

In Table 8 are shown results of the *t* tests computed to evaluate the significance of the differences between pairs of conditions. It can be seen from this table that performance is significantly better with a vocabulary of nine words than of three (cf. B with E, G with H, and I with J). But when the grammar changes from three basic strings to five, R scores remain pretty much the same.

The A-D ratios rise automatically with an increase in the size of the set, as we noted earlier, so that the differences do not reflect a difference in the subjects' performance. Time, except for (I,3) remains constant for the several conditions.

In comparing the minimum sets of (I,3,13) and (I,9,13) with each other, we are impressed with the difference between the two conditions as evident from all three measures in Table 5. The reader will recall that these sets are identical in every respect except for vocabulary; the number of different words, then, does affect the nature of the strategy used. But when we turn to Language (I,9,13) and (I,3,23) which vary in two parameters—in vocabulary and in the number of instances per formula—we notice that the R score differences vanish, that the A-D ratios for (I,3,23) should be (and are) higher by virtue of the size of the set, and that time differences now appear.

The Random group and the (I,3,13) group arrive at similar strategies but at divergent speeds (cf. Columns A and B in Table 8).

Table 9 compares two types of strategies in terms of speed of sorting. Individual data have been clipped for the two extremes of the strategies rated as low and high. Column A is a listing of the amount of time spent per word by subjects who were rated low, that is, who scored either 1 or 2 on their strategies, and Column B presents the data for all subjects who scored 9 or 10. We remind the reader that

TABLE 9

TIME SPENT PER SYLLABLE BY INDIVIDUAL SUBJECTS ON STRATEGIES RATED BY THE R SCORE SCALE AS 1 AND 2, AND 9 AND 10

Condition	A Strategies rated low (1 and 2)	B Strategies rated high (9 and 10)
Random (0,3,16)	1.61 2.49 3.09 1.89 20.09	
Median	2.79	
(I,3,13)	1.89 3.67 4.90 1.57	
Median	4.285	
(I,9,13)		9.03 10.18 3.15 7.56
Median		8.245
(III,3,23)$_c$	2.61	6.32
III,9,23)$_c$		5.44 6.96 7.31
Median		6.96
(III,3,23)$_d$	8.58 1.24 1.54 3.22	6.19
Median	2.38	6.19
(III,9,23)$_d$		4.05 5.41 3.13 4.76 14.96 4.20
Median		4.48
Median—All groups	2.55	6.19

strategies which have been rated low are based upon a counting device of some sort; those rated high, either parallel or are equal to the experimenter's grammar.

Several things seem obvious from Table 9. The poorer strategies go with the smaller vocabulary, not with the larger one (see Column A). In the strategies rated high (see Column B) there is either no subject, or only one subject, per condition under a vocabulary of three words. Also note that a subject spends more than twice as much average time per word in the better strategies as a subject using the less efficient ones.

Discussion

The Number of Different Alternatives (Vocabulary). The results indicate that the size of the vocabulary *does* affect pattern conception. We notice that when nine different words were used the subject was able to decode and classify information more readily than when only three different words were used. Since pattern conception is a matter of seeing meaningful relations among the parts of a set of messages, the function of the number of different alternatives should be to convey information to the decoder about these relations by pointing up similarities and differences among the messages. Information about similarity is usually conveyed through repetition, and about the difference by the use of a different symbol. When the number of alternatives is too small, the encoder is forced to use the same symbol repeatedly even when the messages are different. His problem is initially the same as that of an artist who has to paint a multicolored landscape but who only has one, or maybe two, colors to paint with. When the number of alternatives is small, for example, three as in some of our languages, there is interference in grouping the messages because of the exaggerated similarities among the sequences. With a larger number of alternatives, for example, larger than nine, one can again expect an interference in grouping the messages, but this time the interference will be due to the exaggerated differences among the sequences. This again will tend to obscure the pattern. Perhaps between these two extremes of too few and too many alternatives there is a number which serves the ideal function of emboldening the pattern for the decoder, other things being equal.

Individual data for 3-group (any grammar group tested with three alternatives) show that when faced with haphazard similarities subjects resort to some kind of a counting device. Five subjects out of 30 in the 3-group followed the Length strategy (counting) and 12 subjects in the same group used counting of some other form. Thus a total of 17 subjects out of 30 followed some kind of a counting device. None of the subjects in the 9-group used the Length strategy and only 5 subjects used some shade of the counting strategy. Seventeen subjects in the 3-group as opposed to 5 in the 9-group is not too small a difference to overlook. Subjects in the 9-group, we suggest, can decode and recode the information better, hence their strategies are more closely related to the kind of information that is contained within the set. The strategies of these subjects, therefore, correspond more nearly to the grammar of the language.

In Grammar III we notice that although subjects spend the same amount of average time under all four conditions they do not employ equally good strategies in all the four conditions. Strategies used by the 9-group are more successful than the ones used by the 3-group. Thus once again evidence points to the number of different alternatives as a critical variable.

Now turn to Conditions B and D in Table 7 to make an intralanguage comparison within Grammar I. The results suggest that there is an interaction effect between vocabulary and the number of instances for each recursive loop in a basic string. Although the number of alternatives remains three, performance improves with an increase in the number of instances (see D). With four instances for each loop, subjects perform almost as well as subjects of (I,9,13); they take more time with the set, however. Whether this increase in time for the (I,3,23) group is due to the larger size of the set or because it takes longer for the 3-group to find the pattern than it takes the 9-group, it is hard to say. Since groups (I,3,13) and (I,3,23) take the same amount of time for different sized languages, it could be argued that set size is less critical than the vocabulary factor.

Structural Similiarity (Grammar). The other notable parameter in pattern conception seems to be the structural similarity of sequences. In their studies on visual pattern recognition, Fitts et al. (1956) raised the problem of "different types and degrees of redundancy and of examining the effects of complexity of variables influencing pattern recognition" (p. 9). Although the measure of redundancy does not apply to our materials, as pointed out earlier, the notion of repetitiveness does. By substituting the phrase structural similarity for redundancy in the quotation from Fitts et al. we can state our own problem thus: What is the effect of the type and degree of structural similarity on pattern conception?

Similarities among sequences depend upon two kinds of repetitiveness: the repetition of words because of the limited vocabulary used and repetition due to the grammatical structure of the strings. In the previous section we discussed repetitiveness based upon the number of different alternatives. Structural repetitiveness was defined in terms of the number of basic strings contained in each language and the number and placement of the loops within the basic formulas. A particular grammar selects a particular type of repetitiveness. The problem arising from such selection is: Is the pattern of similarity and difference more difficult in one set than in another?

In scaling the languages for structural similarity, one could begin with the zero condition in grammatical rules, as in the random set of strings, where rules of formation are absent. Results show that in the absence of structural similarities among the strings a subject searches for peripheral similarities and he settles for the simpler strategies such as those of Length or of Initial similarity. Peripheral strategies, however, are not solely the function of nongrammatical structure. We noted earlier, that when the number of alternatives is small, as in (I,3,13), that even when the language has a grammar, the grammar of the set becomes so obscure through word repetition that a subject treats the set as a random one. In such an instance, the strategies used for a grammatical set will approximate those of the random set.

When we turn to the finite state systems, which of course are grammatical in structure, structural similarities consist in the *number* and *kind* of formulas contained within the grammars as explained before. Some questions that now arise are: (*a*) When a system is enlarged by adding one or more parts, do we increase variance or its invariance or both? Thus, are three formulas easier to handle than five? (*b*) Is the quality rather than an increase in the number of parts a critical factor in strategy construction? Are five rules with replication better than five rules without replication? More specifically, is Grammar (III,c) more simple than (I,3) and (III,d) or is it more complex?

The results do not give any clear indication either as to the degree or the directionality of the grammar effects. In the Exploratory Study, the data indicate some improvement in performance with the increase in the number of basic strings, from two strings to four, and to five (see Table 3), but the time also increased correspondingly. The better controlled experiments of the later studies give no clear indication of significant change in performance with an increase in the number of basic strings, the number of states, and the number of rules of transition. The relative difference in the grammar effects based on these factors is hard to assess for two reasons: first, there are interaction effects of other variables with grammar; second, there is no adequate measure of structural similarity. Since grammatical similarity is more structural than peripheral, it can easily elude a subject but unless a subject takes advantage of grammatical repetitiveness implicit in the language, he fails to see structure. How well he will work with the grammar of the language will depend, in part, upon the way the messages have been encoded and, in part, upon how often they have been repeated. Without a sufficient number of different alternatives to indicate similarities and differences among the parts of the system, and without an adequate number of repetitions of the messages, the grammar of the system will remain obscure to a subject and its effect almost negligible upon his recoding behavior. On the other hand, when a subject does figure out the grammatical principle in any one subset, it seems to matter less how many subsets there are in the entire system. It then becomes simply a matter of transfer of a principle. Essentially, we can argue that there is only one basic principle of variance and invariance involved in a finite state system. All the formulas—either one, or three, or five, or any number—are merely variations of this single principle of additions or deletions of symbols to a basic, invariant, sequential pattern. Usually, a subject applies one method of classification fairly consistently to all instances in the set. If this method of classification happens to approximate to the experimenter's grammar, it seems to matter little whether three formulas have been built into the structure of the language or five. And if his strategy is different from the experimenter's grammar, then it matters even less how degenerate the grammar is, because the subject has missed the point of the grammar, anyhow.

We would like to venture another guess to explain the absence of a clear difference in the Grammar groups as measured by the kind of strategies subjects used. Presumably, each time one or more parts is added to a grammatical system, the information function is jointly and simultaneously affected by the increase in complexity and redundancy of the system. In other words, confusion and facilitation go hand in hand in grammatical redundancy and these effects tend to cancel each other. Thus more formulas neither make the task easier nor do they make it more difficult.

Although the quantitative factors of strings, states, and rules of transition seem to make little difference in performance in the present study, recursive loops create some problems for a subject. Two common difficulties pertaining to loops characterize the responses of most of our subjects. These difficulties we shall label as: the zero concept difficulty and the double-loop concept difficulty.

The zero concept: It was not infrequent to find a subject puzzled about the basic string of a formula. For example, if one of the formulas in a grammar was a (c) b b a, its basic string, a $(c)^0$ b b a, was often held as a separate card for a considerable period of time and was even placed in a separate category by some subjects. This observation seems more typical of the double-loop formula to be discussed below. When a subject had decoded a grammatical rule as the experimenter had structured it, he could easily see the arithmetical progression in the development of a loop move from 1 to *n*, even though quite deliberately some exponents in the progression had been left out. But it seemed difficult for a subject to see that the exponent of the loop could also be zero. Viewed historically, the zero concept was not easily come by either. Perhaps it is difficult for one to think of starting a series with nothing. Normal counting frequently begins with 1 and moves up to *n*. It is, therefore, quite possible that this way of counting interferes with the necessity to begin the series with the zero notation in our formulas.

The double-loop concept: The zero concept seemed even more difficult for subjects to handle in the double-loop formulas. This was so because the double-loop in itself was a hard concept for the subject. We have described the double-loop as a formula consisting of two variables, instead of just one, within an invariant sequence of symbols. It has been schematized thus: a (c) b c (b) a.

An R score of 8 or 9 signifies that subjects broke up the double-loop subset into three further sub-subsets as follows: a $(c)^{1-n}$ b c a, a b c $(b)^{1-n}$ a, and a $(c)^{1-n}$ b c $(b)^{1-n}$ a. The basic string was either held as a separate category, or it was placed with one of the first two sub-subsets. In watching subjects at work, it was fairly obvious to the experimenter that subjects were having considerable difficulty with the double-loop subset. They insisted upon applying the simple rule of one loop varying from 0 or 1 to n to this complex formula, and were confused with the similarity within the three sub-subsets. This can be very clearly seen in the thinking-aloud responses of our subject in the Illustrative Study. One could explain this difficulty on the basis of two characteristics of pattern response. The first, that the subject formulates his method early in the game and more or less stays with it; and the second, that the subject seeks to apply one pattern consistently throughout to all parts of the system.

Reiss pointed out that a subject constructs his "model" fairly early, and then begins to stabilize around it as he tests it with successive negative or positive instances. From the thinking-aloud responses of our subjects and on the basis of data from other studies on the evolution of strategies not reported here, Reiss seems to be right. Those subjects who initiated a counting strategy in the beginning showed a strong counting bias, and those who used the initial strategy stuck pretty closely to the left-to-right rule of comparing sequences. Exceptions did occur in the Random group, or when the method of presentation of instances was varied. In the Random set, one could find the subject trying all sorts of strategies, and finally settling on a perceptual one, when nothing else seemed to work. But in the sets where the subject had once perceived the structural order in a subset fairly early, or even later, he did not give it up. This does not mean, of course, that there is no further refinement of the strategy with an increase in the amount of evidence. It simply means that more instances only help to sharpen or modify the "model" already constructed.

This kind of fixation in strategy ties in with the point we are making about consistency in sorting. By and large, most subjects work with one method of classification for all the subsets they make. They do not seem to want to shift gears in their strategy from subset to subset. This characteristic way of responding is independent of the input, and it seems to give us some information about the human channel itself. It is, therefore, important to note that not only does a subject organize the input sequences into categories, but that he does this on the basis of a consistent principle.

Time

Since time did not seem to be a simple function of a single variable we have postponed its discussion; a few remarks are in order now. We note first, that the amount of time taken for the task is not a unique function of the size of the set although size may have something to do with it. Next we note that time seems to be positively correlated with strategy. And finally we see that time is affected by the size of the vocabulary, though not without exception.

Time and Size of Set. First, consider the relation between time and size of set. One is impressed by a lack of orderly variation in the amount of time corresponding to an increase in the size of the set. The group total time spent on a set of 13 instances was 49.56 seconds for the larger and 80.36 seconds for the smaller vocabulary. When the size of the set increases to 16 (larger vocabulary), time reduces to 39.16 seconds, and when the size increases to 18 instances (smaller vocabulary), time goes up to 67.16 seconds. Subjects take more time with a set of 13 instances developed with a vocabulary of three words, than they do with a set of 23 instances irrespective of vocabulary size. The small number of instances does not seem to make the subject's task more manageable, but rather, it seems to make it more difficult, possibly by obscuring the pattern. A small set of instances gives him less evidence upon which to base his induction of the appropriate categories. The variables which interact with the size of the set apparently are far more critical in the time measure than the size (within limits) of the set itself. It is difficult to determine the true time function on the basis of adding just one more instance to the set. It seems one must know what that one more instance is, and also, what some of the other characteristics of the set are.

Time and Strategy. The longer time measures are correlated with the more successful strategies. Even at the risk of explaining the obvious, we shall state our point: complex strategies take more thinking and thinking takes time. We define thinking in this situation as a search for relations among parts of the information received, in order to recode this information into suitable units for use and for storage. The human link in any communication system is responsive not only to the immediate reception and use of information—its function is a far more complex one of storing this information for later use. As a finite system, the organism is confronted with an infinite amount of information from its environment; the quality of its efficiency is on trial on the basis of the recoding scheme the system envisages. How much and how well the old information can be reconstructed from the recoded units will determine the degree of efficiency of the thinking involved. A structural strategy makes the reconstruction of the original message far more probable than a simple counting strategy, for instance. It is, therefore, not difficult to see why it would also take more time. Even Donders (cited in Boring, 1950, pp. 147-149) was suggesting a hundred years ago that the longer reaction times were an indication of complex processes. We might speculate at this point and add that if we could force each subject to work with the information given in the set for a longer period of time (as we forced the subject described in the Illustrative Study) by presenting the information to him one instance at a time, we would expect more subjects, possibly all, to come closer to the original structure of the grammar, despite the complexity of the stimulus conditions. For example, one subject in the (I,3,13) group spent next to the longest time of any and came up with a better solution than any of the other subjects in that group. Give a man sufficient time and he will find some way to reduce the heterogeneity of his environment to lawful change.

CONCLUDING OBSERVATIONS

It is clear from the data that pattern conception is a characteristic way in which people respond to a set of linguistic sequences, irrespective of the structure of the sequences. In an "open system" such as ours (unstructured in instruction and in meeting any given criteria) the subject must define the problem himself, construct some sort of working hypothesis to deal with the information he has extracted from the input, and look for evidence to see if his hypothesis holds. He then discovers gaps in the information he has acquired, he searches for new evidence to fill the gaps, and finally he accepts his plan or rejects it if the gaps remain unfilled. Usually, though a subject may show intolerance of uncertainty, he may occasionally decide to live with his problem. The strategy or plan he settles on is partially a function of some experimental variables.

It was interesting to observe that reminiscent of the Kantian categorical imperative, our subjects seemed to be bound by a categorizing imperative. And what is more, their categorizing seems to fall within a limited range. Whether there were 13 instances in a set, or 50, and whether there were 15 subjects tested or 150, it did not seem to matter; the mean number of categories ranged between five and seven (see Figure 4). By some strange coincidence, the integer that has been "persecuting" Miller (1956) for a decade, turns up once again in our data. What does one make of it?

Figures 4, 5, and 6 all relate to the number of categories subjects make. The function in Figure 4 represents data for the mean number of categories subjects made for the different sized sets. Data show that the number of categories increases somewhat with an increase in the total number of instances in the set, but that this increase is proportionately less.

The average frequency of the different types of categories subjects make is presented in Figure 5. By the type of category is meant the number of instances subjects put into one category; for example, there may be a one-instance category, a category with two instances in it, with three, and so on. The number of instances in a category and the total number of categories are interdependent factors, of course. Either a subject prefers to make a certain number of categories, or he prefers to put a certain number of instances together in one category. The more frequent categories are those with two, three, or four instances per category.

Figure 6 treats the same data in a slightly different way. It presents individual cumulative curves for different language conditions to show the probability that a subject will make at least *x* categories. Note that for all conditions 50% of the subjects average between five and six categories.

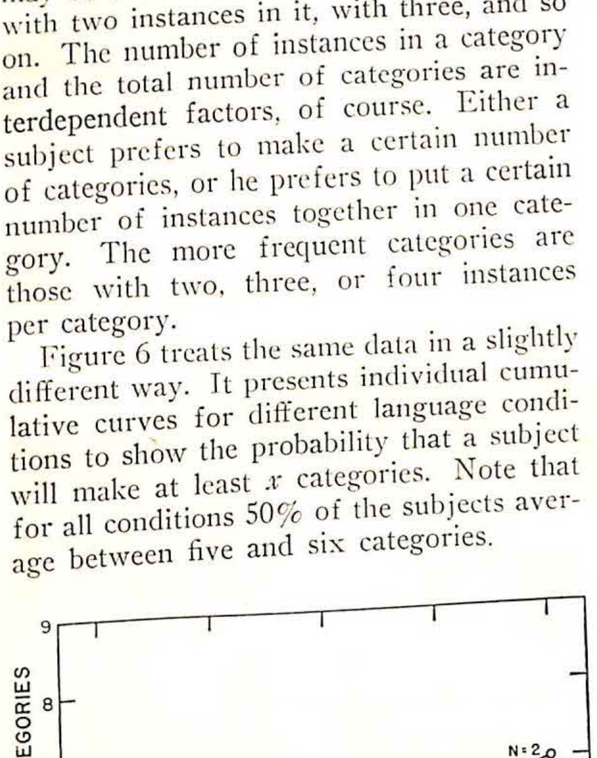


FIG. 4. Mean number of categories per size of set. (*N* denotes the number of subjects in each group. The single filled circle represents the Random group.)

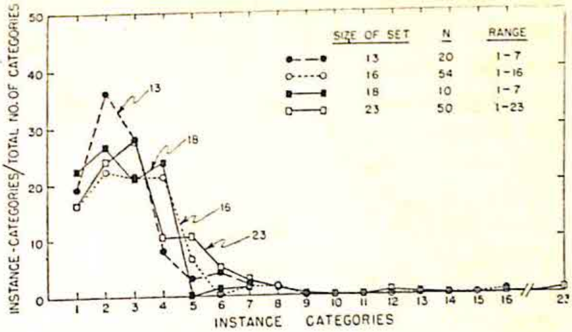


FIG. 5. Mean number of instance categories per size of set. (Instance category refers to the number of instances placed in a category. The *y* axis represents what percentage an instance category is of the total number of categories made in each language. For example, in the 23-instance languages, 50 subjects made a total of 338 categories, of which 54 categories were one-instance categories.)

We have stated above that although the number of instances in a set has something to do with the number of categories subjects make, it is not a determining factor. The relation of the number of basic strings in a set to the number of categories seems even less important, except when a subject sees the grammatical pattern accurately. Quite frequently subjects wished to know how many categories they were expected to make. When they were informed that the choice of number rested with them, they seemed to guide their sorting by a desire not to have "too many" categories. This was especially true of subjects employing some type of formula strategies. "It's a shame to have a one-card category" was not unusual for a subject to say when an instance remained unclassified. If, however, the choice before the subject was between consistency in sorting and the number of categories, he preferred consistency to number.

Seven, then, is still a magic number. The "number seven, plus or minus two" denotes according to Miller (1956), "the span of absolute judgment" and the "span of memory" (p. 90). We interpret the spans to mean that there is a finite number of isolated events, or of groups of events, within the neighborhood of seven, which the human link in a communication system

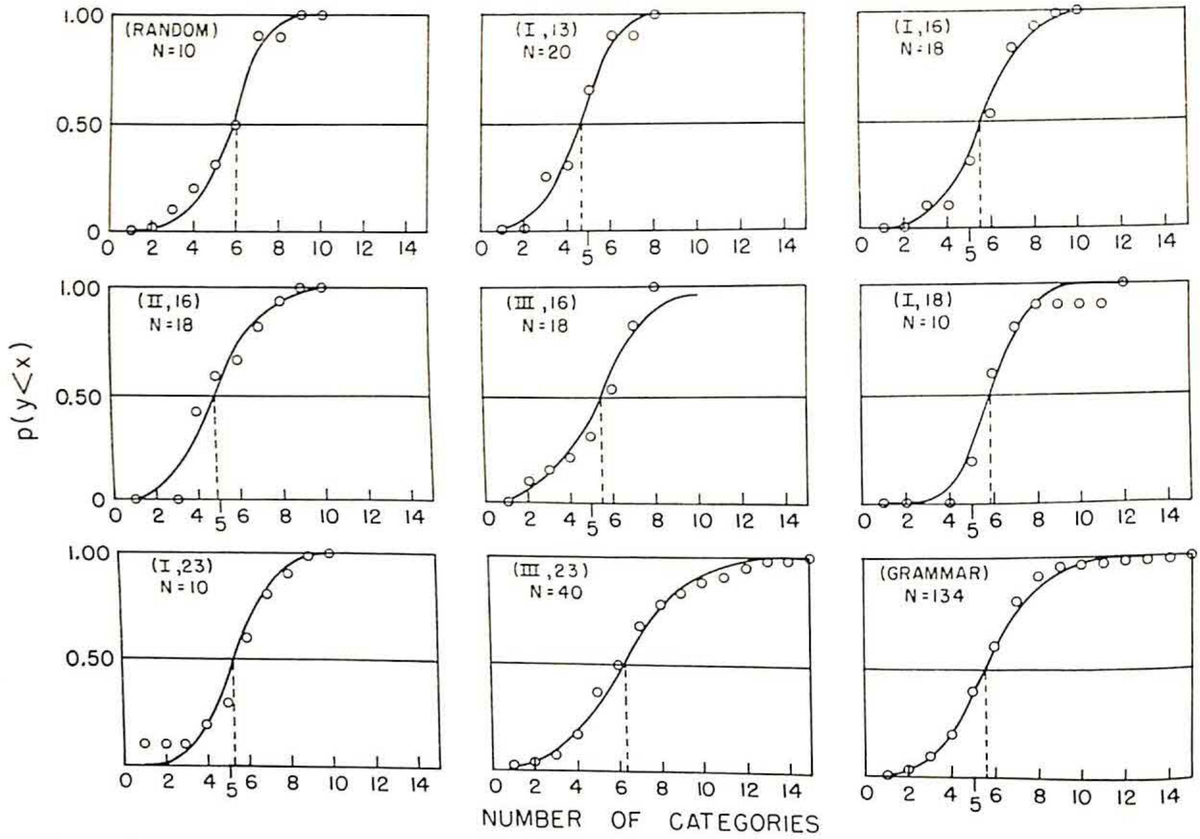


FIG. 6. The probability that a subject will make at least x categories is represented by $p(y < x)$ for nine groups of subjects. (The S-shaped functions represent theoretical curves fitted to the data on probability paper for each group separately. Circles represent actual data. Data from all Grammar groups have been combined in the last figure where $N = 134$.)

can best handle. It seems, then, that in dealing with a bulk of information the human channel is constrained by its "channel capacity," to use Miller's terminology, to process the given information within the given range of categories. It seems as though, because of the limits imposed by the memory span upon pattern conception, one might begin to develop a notion of conceptual span as well. Not only is memory limited by the number of psychological units it can handle, but pattern conception must function under the same limitation.

Although the span of immediate memory is limited by the number of items, the amount of information the items contain can be increased by a process of recoding. Pattern conception is a description of the recoding process itself. The linguistic recoding involved in the type of studies presented above is basic to the understanding of memory and thinking. The study is in its infancy; much more research is warranted.

SUMMARY

It is advantageous for an organism, for both communication and storage of information, to organize and symbolize the stimulus energies in its environment. Perceptual organization is one way to reduce the apparent variability of the environment, conceptual symbolization is another. Together, these two forms of information processing provide an organism with powerful weapons to deal efficiently with the heterogeneity of the environment. Numerous psychological studies have indicated that perceptual recoding is limited by the amount of information given; is conceptual recoding similarly related in any specifiable way to the given information?

Conceptual recoding is here discussed in terms of the discovery and formulation of rules to describe cognitive structures. Language provides one of the best illustrations; verbal recoding of experience seems to be the essence of human thinking. It is to language that one can turn, therefore, to find the materials needed for experimental study of the processes of conceptual recoding. Natural languages, however, are tremendously complicated; some of the same processes can be studied more simply with artificial language systems. In the present study finite state systems were used to simulate natural languages. They are simpler, their formal structure can be stated precisely, yet they provide the kind of sequential patterning that is characteristic of all verbal communication. Subjects were given a set of linguistic strings, random or constrained, generated by some finite state systems; they were asked to group the strings in some appropriate way. In order to perform such a task all subjects employ some intelligible, recognizable plan for sorting the strings. The number of different strategies that are observed to occur is small and subjects are easily classified according to the strategy they adopt. Some strategies are more complicated than others; they range from simple, ready-made rules for counting and alphabetizing to ad hoc inventions of rules that a subject considers implicit in the information he has been given. The former are more perceptual, the latter more conceptual in spirit, but both are ways of describing the structural rules of the system.

Pattern conception depends upon the kind and amount of the evidence supplied by the experimenter. In the present studies, the length of the messages, the number of instances, the number of alternatives, and structural similarity are some of the experimental variables by which messages were varied. The results show that the variety of symbols (number of different alternatives) with which the messages are constructed is far more critical in controlling pattern conception than is the length or the structure of the messages. Two sets of messages, identical in length, in rules of formation, and in the instruction given subjects, but differing in the number of symbols, elicit qualitatively different responses. A large variety of symbols tends to elicit the invented strategies whereas the small variety generally leads to ready-made rules. This result indicates that the subject is able to recode the given information more effectively when there is a variety in the symbols from which it is composed.

A parallel advantage is observed by increasing the amount of evidence provided to a subject for a given rule. The number of instances per rule was varied from two to three to four instances, while other factors of vocabulary, grammar, and instructions were held constant. The strategies that subjects used under each of these conditions were evaluated in terms of two statistical techniques developed specifically for the purpose of strategy analysis. As the amount of evidence increased, the strategies improved. A larger quantity of evidence facilitates recoding of the sequences in terms of variant and invariant symbols.

Earlier experiments on perception have indicated that redundancy may be of several kinds, and that the influence of redundancy upon responses varies with its quality. Although we have not applied an information measure of redundancy to our materials (the finite state systems that we used are too simple to require such statistical treatment) we have worked with a related conception of structural similarity (rules of formation) in linguistic sequences. Languages varying in number and complexity of rules of formation, but similar in length and size of vocabulary, do not yield markedly different results. The fewer rules embedded in a language do not necessarily make that language simpler. Data indicate that structural redundancy is less critical in pattern conception than is the variety in symbols or the amount of evidence. More experiments with better controls on redundancy are needed to determine the effect of this variable on pattern conception.

Two specific difficulties were found to characterize the responses of many subjects in almost all of the experimental groups. These difficulties were the concept of zero and the concept of double variation in a linguistic sequence of variant and invariant symbols. Subjects could delete or add the variable from 1 to *n* fairly easily, but they struggled with the string when the exponent of the variable was equal to zero. In the case of more than one variable in a string subjects tended to group its instances into as many subsets as there were variables in the string. This multiple-loop difficulty is not unrelated to the zero difficulty, however.

Pattern conception, apparently, is not a simple function of any one experimental variable. This point was most clearly evident from the time measures, which did not correlate consistently with any one experimental condition. It was reasonably clear, however, that the longer time measures correlated positively with the better strategies. Presumably, the central processes of recoding and thinking take time, which causes the delay in response. There was little evidence for an increase in time with the increasing length of the set.

Finally, all studies led us to conclude that, irrespective of the type of finite state system used—its structure, vocabulary, length, or anything else—the average number of categories that subjects make is five or six. Although we were not looking for this point, this result is in keeping with the findings from some studies on "channel capacity," and it indicates that thinking like memory is centrally affected by the channel capacity of the organism.

REFERENCES

Aborn, M., & Rubenstein, H. Information theory and immediate recall. *J. exp. Psychol.*, 1952, **44**, 260–266.

Adams, O. S., Fitts, F. M., Rappaport, M., & Weinstein, M. Relations among some measures of pattern discriminability. *J. exp. Psychol.*, 1954, **48**, 81–84.

Anderson, Nancy S., & Leonard, J. A. The recognition, naming, and reconstruction of visual figures as a function of contour redundancy. *J. exp. Psychol.*, 1958, **56**, 262–270.

Attneave, F. Some informational aspects of visual perception. *Psychol. Rev.*, 1954, **61**, 183–194.

Attneave, F. Symmetry, information and memory for patterns. *Amer. J. Psychol.*, 1955, **68**, 209–222.

Attneave, F. *Applications of information theory to psychology: A summary of basic concepts, methods, and results.* New York: Holt, 1959.

Bartlett, F. *Thinking: An experimental and social study.* New York: Basic Books, 1958.

Boring, E. G. *A history of experimental psychology.* New York: Appleton-Century-Crofts, 1950.

Brogden, W. J., & Schmidt, R. E. Acquisition of a 24-unit verbal maze as a function of the number of alternative choices per unit. *J. exp. Psychol.*, 1954, **48**, 335–338. (a)

Brogden, W. J., & Schmidt, R. E. The effect of number of choices per unit of a verbal maze on learning and serial position errors. *J. exp. Psychol.*, 1954, **47**, 235–240. (b)

Bruner, J. S. Goodnow, J. J., & Austin, G. A. *A study of thinking.* New York: Wiley, 1956.

Carmichael, L., Hogan, H. P., & Walter, A. A. An experimental study of the effect of language on the reproduction of visually perceived forms. *J. exp. psychol.*, 1932, **15**, 73–86.

Chomsky, N. *Syntactic structures.* Hague: Mouton, 1957.

Chomsky, N., & Miller, G. A. Finite state languages. *Inform. Control*, 1958, **1**, 91–112.

Fitts, P. M., Weinstein, M., Rappaport, M., Anderson, Nancy S., & Leonard, J. A. Stimulus correlates of visual pattern recognition: A probability approach. *J. exp. Psychol.*, 1956, **51**, 1–11.

French, R. S. The discrimination of dot patterns as a function of number and average separation of dots. *J. exp. Psychol.*, 1953, **46**, 1–9.

French, R. S. The effect of instruction on the length-difficulty relationship for a task involving sequential dependency. *J. exp. Psychol.*, 1954, **48**, 89–97. (a)

French, R. S. Identification of dot patterns from memory as a function of complexity. *J. exp. Psychol.*, 1954, **47**, 22–26. (b)

French, R. S. Pattern recognition in the presence of visual noise. *J. exp. Psychol.*, 1954, **47**, 27–31. (c)

Goldstein, K., & Scheerer, M. Abstract and concrete behavior: An experimental study with special tests. *Psychol. Monogr.*, 1941, **53**(2, Whole No. 239).

Hanfmann, Eugenia, & Kasanin, J. A method for study of concept formation. *J. Psychol.*, 1937, **3**, 521–540.

Heidbreder, Edna. The attainment of concepts: II. The problem. *J. gen. Psychol.*, 1946, **35**, 191–223.

Hovland, C. I. The generalization of conditioned responses: I. The sensory generalization of conditioned responses with varying frequencies of tone. *J. gen. Psychol.*, 1937, **17**, 125–148. (a)

Hovland, C. I. The generalization of conditioned responses: II. The sensory generalization of conditioned responses with varying intensities of tone. *J. genet. Psychol.*, 1937, **51**, 279–291. (b)

Hovland, C. I. A "communication analysis" of concept learning. *Psychol. Rev.*, 1952, **59**, 461–472.

Hull, C. L. Quantitative aspects of the evolution of concepts. *Psychol. Monogr.*, 1920, **28**(1, Whole No. 123).

Johnson, D. M. *The psychology of thought and judgment.* New York: Harper, 1955.

Koffka, K. *Principles of gestalt psychology.* New York: Harcourt, Brace, 1935.

Köhler, W. *Gestalt psychology.* New York: Liveright, 1929.

Köhler, W. *The place of value in a world of facts.* New York: Liveright, 1938.

Köhler, W. *Gestalt psychology: An introduction to new concepts in modern psychology.* New York: Liveright, 1947.

Krechevsky, I. "Hypotheses" vs. "chance" in the pre-solution period in sensory discrimination-learning. *U. Calif. Publ. Psychol.*, 1932, **6**, 27–44.

Luce, D. R. A survey of the theory of selective information and some of its behavioral applications. New York: Bureau of Applied Social Research, 1954. (Behavioral Models Project (NR 042-115) CU-23-54-NONR-266621-BASR).

Miller, G. A. *Language and communication.* New York: McGraw-Hill, 1951.

Miller, G. A. What is information measurement? *Amer. Psychologist*, 1953, **8**, 3–11.

Miller, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychol. Rev.*, 1956, **63,** 81–97.

Miller, G. A. Free recall of redundant strings of letters. *J. exp. Psychol.*, 1958, **56,** 485–491.

Miller, G. A., & Chomsky, N. Pattern conception. Paper delivered at Conference on Pattern Detection at University of Michigan, 1958.

Miller, G. A., & Selfridge, Jennifer A. Verbal context and the recall of meaningful material. *Amer. J. Psychol.*, 1950, **63,** 175–185.

Mosteller, F., & Bush, R. R. Selected quantitative techniques. In G. Lindzey (Ed.), *Handbook of social psychology.* Vol. I. New York: Addison-Wesley, 1954. Ch. 8.

Osgood, C. E. *Methods and theory in experimental psychology.* New York: Oxford Univer. Press, 1956.

Quastler, H. (Ed.) *Information theory in psychology: Problems and methods.* Glencoe: Free Press, 1955.

Razran, G. H. S. Studies in configurational conditioning: V. Generalization and transposition. *J. genet. Psychol.*, 1940, **56,** 3–11.

Razran, G. H. S. Stimulus generalization of conditioned responses. *Psychol. Bull.*, 1949, **46,** 337–365

Reiss, D. Subjective models of a simple finite-state grammar: A study of their attainment and characteristics. Unpublished honors thesis, Harvard University, 1958.

Rubenstein, H., & Aborn, M. Immediate recall as a function of degree of organization and length of study period. *J. exp. Psychol.*, 1954, **48,** 146–152.

Shannon, C. E., & Weaver, L. W. *The mathematical theory of communication.* Urbana: Univer. Illinois Press, 1949.

Smoke, K. L. An objective study of concept formation. *Psychol. Monogr.*, 1932, **42** (4, Whole No. 191).

Spence, K. W. The solution of multiple choice problems by chimpanzees. *Comp. psychol. Monogr.*, 1939, **15,** No. 3.

Thorndike, E. L. *Human learning.* New York: Century, 1931.

von Ehrenfels, C. On gestalt qualities. *Psychol. Rev.*, 1937, **44,** 521-524.

Wertheimer, M. *Productive thinking.* New York: Harper, 1945.

Witkin, H. A. I. "Hypotheses" in rats: An experimental critique. *J. comp. Psychol.*, 1940, **30,** 457–482.

Woodworth, R. S. *Contemporary schools of psychology.* New York: Ronald, 1948.

(Received April 11, 1960)

APPENDIX A

Agreement-Disagreement Ratio

The Agreement-Disagreement ratio is a statistic devised to measure the amount of correlation between sortings made by any two persons. The technique of calculating the amount of agreement between different sortings is as follows:

Consider the sortings made by any two subjects X and Y:

X—(1 2 3 4 5) (6 12 16) (9 14)
(7 8 10 11 13 15)

Y—(1 2 4 5) (3) (6 10 12 15)
(7 8 9 11 13 14) (16)

Step 1. Construct sets from the intersection of these two classifications:

Intersection (1 2 4 5) (6 12) (7 8 11 13)
(9 14) (10 15) (3) (16)

The intersection is constructed in such a way that every set in the data for X and Y can be reconstructed by combining sets from the intersection, and yet the sets in the intersection are as large as possible. Otherwise said, every distinction among the strings that either X or Y indicated is preserved in the intersection.

Step 2. Calculate the number of agreements for each set. For example, if a set contains four strings there are 3 + 2 + 1 = 6 pairs of strings that agree with one another in the sense that they are placed in the same class and considered to be equivalent. Below are tabulated the number of pairs of strings that agree as a function of the size of set.

size	*number agreements*
1	0
2	1
3	3
4	6
5	10
x	$x(x-1)/2$

Thus, for example, a set of 11 strings would yield 55 pairs that were in agreement. If size is x then the number of agreements is given by the equation $A = x(x-1)/2$. So the second step is to add up the number of agreements in this sense for X, Y, and the intersection:

subject X		*subject Y*		*intersection*	
5	10	4	6	4	6
3	3	1	0	1	0
2	1	4	6	2	1
6	15	6	15	4	6
	29	1	0	2	1
			27	2	1
				1	0
					15

Step 3. Calculate the A-D ratio from these three numbers, 29, 27, 15, by substituting them into the following equation:

$$r = 1 - \frac{x + y - 2\ \text{Intersection}}{60}$$

$$r = 1 - \frac{29 + 27 - 2\ (15)}{60} = 1 - \frac{26}{60}$$

$$= 1 - .443 = .567$$

APPENDIX B

RATING SCALE FOR SUBJECTS' STRATEGIES

The second column gives the average scores for the ratings assigned to 20 strategies by five judges. The first column rounds off the average scores for purposes of establishing a 10-point rating scale.

Final Rating Score	Average Score[a]	Kind of Strategy
1	1.0	General Size
	1.5	Length
2	2.4	Initial$_1$, or Middle$_1$, or Final$_1$
	2.5	Alphabet
3	2.7	One striking word with Initial or Final one
	3.1	Initial$_{(2 \text{ or } n)}$, Middle$_{(2 \text{ or } n)}$, Final$_{(2 \text{ or } n)}$
	3.1	Initial$_1$-Final$_1$
	3.4	Length plus Alphabet
	3.4	Initial$_1$ plus Alphabet or Initial plus Length
	3.6	Rhythm
4	3.8	Mixed methods for entire set
	3.8	Mixed methods by groups
	4.3	Alternating word or complex counting
5	5.3	Symmetry and Rhythm
	5.6	Initial plus complex substitution or ordering
6	6.2	Initial$_{(92 \text{ or } 3)}$-Final$_{(2 \text{ or } 3)}$
7	7.4	Maximum Identity
8	8.2	Initial-Middle-Final
9	9.0	Less than basic string (breaking up the formula by loops)
10	10.0	Formula (same as experimenter's grammar)

---

[a] The steps between scores are not necessarily equal.